FOUNDED BY

W. F. R. WELDON, F.R.S., FRANCIS GALTON and KARL PEARSON

EDITORIAL EDITOR

E. S. PEARSON

ASSOCIATE EDITORS

M. G. KENDALL JOHN WISHART

In consultation with

HARALD CRAMÉR J. B. S. HALDANE

F. N. DAVID H. O. HARTLEY

E. C. GEAR D. G. KENDALL

Re-issued photo litho 1966

ISSUED BY

THE BIOMETRIKA OFFICE, UNIVERSITY COLLEGE, LONDON

PRINTED AT THE UNIVERSITY PRESS, CAMBRIDGE

(*Issued 8 December 1966*)

VOLUME 42, PARTS 3 AND 4 DECEMBER 1955

POPULATION ESTIMATION BASED ON CHANGE OF COMPOSITION CAUSED BY A SELECTIVE REMOVAL

BY DOUGLAS G. CHAPMAN*

Department of Mathematics, University of Washington and Readership in the Design and Analysis of Scientific Experiment, Oxford University

1. INTRODUCTION

It has been noted in wildlife studies that it is possible to estimate the size of an animal population on the basis of the change of sex ratio following a kill concentrated on one sex. It is difficult to ascertain to whom this idea should be credited, and what name should be given to it. Scattergood (1954), in a general survey, lists several references to the method. Chapman (1954) called the procedure the dichotomy method of estimation and showed that the usual 'intuitive' estimates are also maximum-likelihood estimates. It is not clear that this title is sufficiently descriptive, and in any case the classification does not necessarily need to be dichotomous. On the other hand, change of composition does seem to reflect the essential aspect of the procedure.

The assumptions involved in this method appear to be simpler and more likely to be fulfilled than those underlying other population estimation procedures. Moreover, it will be shown that non-fulfilment of some of the assumptions may involve no serious consequences. This is one of the phases of the method studied in this paper.

To be balanced against this is the fact that the amount of information yielded is smaller for the same effort than is obtainable from the more usual method of estimation based on the recapture of marked members of the population. A quantitative comparison is given below of the two methods. Some problems of optimum design in the sampling that accompanies a change of composition estimate are considered and some extensions of the procedure studied.

2. BINOMIAL MODEL

The simplest situation arises if a 'closed' population is assumed, i.e. one where there is no emigration or immigration and births and deaths are negligible over the period of the experiment—except for the removal process. As to the removal process, all that is required of it is that the final or cumulative total be known and a breakdown into at least two classes or components. The third assumption for the simple binomial model is that random samples have been taken of the population before and after the removal process. The randomness is in respect to the specified classes or components.

The following notation is needed:

$$\text{Unknown parameters:} \begin{cases} N_i = \text{population size at time } t_i \ (i = 0, 1) \text{ made up of two classes } X \text{ and } Y, \\ X_i, Y_i = \text{size of classes } X \text{ and } Y \text{ at times } t_i. \end{cases}$$

* Research initiated under sponsorship of the Office of Naval Research. The author is currently a Fellow of the Guggenheim Foundation.

It is convenient to write $P_i = X_i/N_i$ $(i = 0, 1)$:

Known parameters:
$$\begin{cases} n_i = \text{size of random samples taken at times } t_i, \\ R_x = X_1 - X_0 = \text{removal from class } X, \\ R_y = Y_1 - Y_0 = \text{removal from class } Y, \\ R = R_x + R_y = \text{total removal.} \end{cases}$$

Random variables: x_i, y_i = the number of classes X and Y respectively in sample i $(i = 0, 1)$.

Analogous to P_i, we define $p_i = x_i/n_i$ $(i = 0. 1)$.

As noted in Chapman (1954) the maximum-likelihood estimates of X_0 and N_0 are

$$\hat{X}_0 = \frac{x_0(n_1R_x - x_1R)}{n_1x_0 - n_0x_1} = \frac{p_0(R_x - p_1R)}{p_0 - p_1}, \tag{1}$$

$$\hat{N}_0 = \frac{n_0(n_1R_x - x_1R)}{n_1x_0 - n_0x_1} = \frac{R_x - p_1R}{p_0 - p_1}, \tag{2}$$

while the asymptotic variances of X_0 and N_0 are

$$\sigma^2(\hat{X}_0) = \frac{\dfrac{X_0Y_0}{n_0}P_1^2 + \dfrac{X_1Y_1}{n_1}P_0^2}{(P_0 - P_1)^2}, \tag{3}$$

$$\sigma^2(\hat{N}_0) = \frac{\dfrac{X_0Y_0}{n_0} + \dfrac{X_1Y_1}{n_1}}{(P_0 - P_1)^2}. \tag{4}$$

In the usage made of this method in wildlife studies it is N_0 that is ordinarily estimated, based as noted above, on the change of sex ratio. The sports kill records show the removal of each sex. In many cases the kill is confined to males; this represents a very favourable situation. It should be pointed out that under the assumptions made it is immaterial whether attention is focused on time t_0 or time t_1.

In the possible application of this procedure to the estimation of fish populations, it must be recognized the removal is not always selective by sex. However, here the classification may be by size or by species. In many fisheries there is a size restriction on fish that may be retained by the fishery. If the biologist can sample in a manner to include randomly other sizes, then this change of composition estimation procedure may be used. In this case, or in case the classification is by species, the experimenter will usually be more interested in estimating X_0.

It may be suggested that where X and Y represent qualitatively different populations, one should expect different catchabilities and consequently the procedure based on such a classification would be unsatisfactory. However, in the rather important case that $R_y = 0$, the estimate remains asymptotically unbiased even if the catchability of the X population is not the same as that of the Y population (for the experimenter or sampler). We repeat that the catchabilities associated with the removal process are irrelevant.

For consider that X and Y have catchabilities λ_x and λ_y and assume the total sample catch has a Poisson distribution. Then given n_i, the sample catch size, x_i, has a binomial distribution with parameters n_i, $X_i/(N_i - Y_i\delta)$, where $\delta = 1 - \lambda_y/\lambda_x$.

Then replacing the random variables in $\hat{X}_0$ by their expectations (which they will converge to in probability as the sample sizes are increased), we have

$$\hat{X}_0 = \frac{X_0(N_0R_x - X_0R - \delta Y_0R_x + \delta R_xR_y)}{(N_0R_x - X_0R - \delta Y_0R_x + \delta X_0R_y)}, \tag{5}$$

and the result stated follows immediately.

However, $\hat{N}_0$ does not have the same desirable property, for replacing x_i, y_i by their expectation under the model of unequal catchabilities

$$\hat{N}_0 = (N_0 - Y_0\delta)\frac{(N_0R_x - X_0R - \delta Y_0R_x + \delta R_xR_y)}{(N_0R_x - X_0R - \delta Y_0R_x + \delta X_0R_y)}. \tag{6}$$

Nevertheless, the fact that X_0 is still estimable indicates that it may be possible to use a quite different species in this estimation procedure, e.g. a scrap fish may form the Y class while the desirable sports fish is the X class. It should be pointed out that the distribution of the estimate will be modified, in case $\lambda_x \neq \lambda_y$, even though $R_y = 0$. In particular, in this case the asymptotic variance of $\hat{X}_0$ is

$$\sigma^2(\hat{X}_0(\delta)) = \frac{1-\delta}{(P_0 - P_1)^2}\left[\frac{X_0Y_0}{n_0}\frac{P_1^2}{(1-P_0\delta)^2} + \frac{X_1Y_1}{n_1}\frac{P_0^2}{(1-P_1\delta)^2}\right]. \tag{7}$$

The variation of $\sigma^2(\hat{X}_0)$ with respect to δ depends on all of the unknown parameters. However, it is possible to state some useful qualitative results. These can be determined simply from a study of the function

$$f(\delta) = (1-\delta)/(1-P_i\delta)^2 \quad (\delta < 1),$$

which attains its maximum at $\delta = 2 - 1/P_i$. This maximum is $[4P_i(1-P_i)]^{-1} \geqslant 1$.

The qualitative results are summarized in the following table:

	$\delta > 0$ $\lambda_x > \lambda_y$	$\delta < 0$ $\lambda_x < \lambda_y$
$P_0, P_1 < \frac{1}{2}$	$\sigma^2(\hat{X}_0(\delta)) < \sigma^2(\hat{X}_0(0))$	$\sigma^2(\hat{X}_0(\delta)) > \sigma^2(\hat{X}_0(0))$
$P_0, P_1 > \frac{1}{2}$	$\sigma^2(\hat{X}_0(\delta)) > \sigma^2(\hat{X}_0(0))$	$\sigma^2(\hat{X}_0(\delta)) < \sigma^2(\hat{X}_0(0))$

The favourable cases occur in the main diagonal; they are favourable in the sense that if $\sigma^2(\hat{X}_0(0))$ is used to estimate the variance or determine a confidence interval, any errors incurred will be on the conservative side. In terms of a working rule for the sample considered, if the undesirable fish are abundant relative to the desirable fish, then the procedure is more satisfactory if the sampling device is such that the desirable fish are more catchable. Since $f(\delta)$ achieves a maximum value of unity at $\delta = 0$ for $P_i = \frac{1}{2}$, it seems reasonable to suggest that when the P_i straddle $\frac{1}{2}$ and do not differ greatly from it, $\sigma^2(\hat{X}_0(\delta))$ will differ little from $\sigma^2(\hat{X}_0(0))$.

One other possibility of error in this procedure arises from the fact that R_x, R_y and R may not be true totals removed from the population. In general, it is to be expected that the removal will be under-evaluated because of the unreported kill as well as the unknown natural mortality.

Suppose that the actual removals are $R_x+\Delta_x$, $R_y+\Delta_y$, $R+\Delta$, where $\Delta_x+\Delta_y=\Delta$ and where we may write $\Delta_x = KR_x$, $\Delta = KR+\epsilon$.

Then for $n_i \to \infty$, the estimate $\hat{N}_0$ converges in probability to

$$N_0\left(\frac{N_0R_x - X_0R - \epsilon R_x}{(N_0R_x - X_0R)(1+K) - \epsilon X_0}\right). \tag{8}$$

It is convenient to denote the term in the brackets of (8) by $1+b$. X_0 also converges in probability, under these conditions, to $X_0(1+b)$.

Consider first the case where $\epsilon = 0$, i.e. the unknown removal of X and Y is in the same proportion as the known removal. Then

$$1+b = (1+K)^{-1} \approx 1-K \text{ for } K \text{ small};$$

hence the relative asymptotic bias is negative and of the order of the unknown proportion of the removal.

For the more important case where ϵ is not zero we restrict attention to the situation where $R_y = 0$. Then

$$1+b = \frac{Y_0 - \epsilon}{Y_0(1+K) - \dfrac{\epsilon X_0}{R}}.$$

If the mortality or unknown removal of the X's and Y's is proportional to the size of these groups, then

$$\Delta_y = Y_0\Delta_x/X_0 = Y_0KR/X_0$$

and

$$\Delta = KRN_0/X_0, \quad \epsilon = Y_0KR/X_0.$$

Hence

$$b = -KR/X_0.$$

In this situation the relative bias is still negative and of an even smaller order than in the case where $\epsilon = 0$.

In the usage of (4) (or (3)) to estimate the large sample variances, in a situation where part of the removal is unknown to the experimenter, errors will arise in the incorrect estimation of the X_i, Y_i. It is easily seen that, while on the average, for sufficiently large samples, X_0, Y_0 are underestimated, X_1 and Y_1 will be overestimated and consequently the estimation errors will compensate. From a qualitative point of view, therefore, it appears that the errors due to bias in the estimation of the variances are of much smaller consequence than the sampling errors.

The cases considered above are by no means the only possibilities that might occur with a portion of the removal unknown to the experimenter. They represent, however, two cases of practical importance. Of course if no assumptions whatsoever are made as to the size and relative magnitudes of Δ_x and Δ_y, it is clear that they may be such as to vitiate the estimation procedure.

3. Optimum sample allocation

Since this method of population estimation involves two samples over which the experimenter has some degree of control, it is natural to inquire as to the optimum disposition of effort between these samples. If N_0 is to be estimated with $\sigma^2(\hat{N}_0)$ to be minimized, subject to the restriction that n_1+n_0 be fixed, then elementary calculus leads to the following equation for the ratio of n_1 to n_0:

$$\frac{n_1}{n_0} = \left(\frac{X_1Y_1}{X_0Y_0}\right)^{\frac{1}{2}}. \tag{9}$$

Formula (9) involves parameters which are unknown at the outset of the experiment. It is possible to use it to suggest qualitative rules of procedure. Consider the most favourable case where the removal is restricted entirely to one class, say the X class, i.e. $R_y = 0$. Then

$$n_1 = n_0\left(1-\frac{R}{X_0}\right)^{\frac{1}{2}};$$

the ratio n_1/n_0 is given in Table 1 for various values of R/X_0, the proportion of the X class removed. Since the removal will rarely exceed 50 % of the removed class, this table suggests that if N_0 alone is to be estimated, n_1 should be chosen only slightly smaller than n_0.

Table 1. *Optimum sample ratio n_1/n_0 for the estimation of N in case there is no removal in the Y class*

R/X_0	0·5	0·4	0·3	0·2	0·1	0·05	0·02	0·01
n_1/n_0	0·71	0·77	0·84	0·89	0·95	0·975	0·990	0·995

If X_0 is to be estimated, the situation is slightly more complicated, but a qualitative rule is possible. Recalling $\sigma^2(X_0)$ from formula (3), this is minimized with n_0+n_1 fixed if

$$\frac{n_1}{n_0} = \left(\frac{X_1Y_1}{X_0Y_0}\right)^{\frac{1}{2}}\frac{P_0}{P_1} = \frac{N_1}{N_0}\left(\frac{X_0Y_1}{X_1Y_0}\right)^{\frac{1}{2}}. \tag{10}$$

Again, consider the case $Y_0 = Y_1$ and hence $R_x = R$. Then the optimum choice of n_0 and n_1 is determined by

$$\frac{n_1}{n_0} = \left(1-\frac{R}{N_0}\right)\left(1-\frac{R}{X_0}\right)^{-\frac{1}{2}},$$

which is shown in Table 2.

Table 2. *Optimum sample ratio n_1/n_0 for the estimate of X in case there is no removal in the Y class*

$P_0=X_0/N_0$	$R/N_0=r$						
	0·25	0·20	0·15	0·10	0·05	0·02	0·01
0·1	—	—	—	—	1·344	1·096	1·043
0·2	—	—	1·700	1·273	1·097	1·033	1·015
0·3	1·838	1·386	1·202	1·102	1·041	1·014	1·007
0·4	1·225	1·132	1·075	1·039	1·016	1·005	1·003
0·5	1·061	1·032	1·016	1·007	1·001	1·000	1·000
0·6	0·982	0·980	0·982	0·986	0·993	0·997	0·998
0·7	0·935	0·947	0·959	0·972	0·985	0·994	0·997
0·8	0·905	0·924	0·943	0·963	0·981	0·993	0·996
0·9	0·882	0·907	0·931	0·954	0·977	0·991	0·996

Table 2 suggests that in a few extreme cases the second sample should be considerably larger than the first for an optimum estimate of X_0; for most cases, however, the samples should be chosen almost equal in size. Since often the biologist will wish to know both X_0 and N_0, it would seem that choosing $n_0 = n_1$ represents a near optimum allocation of effort.

4. Comparison of capture-recapture and change of composition estimation procedures

Rather than merely sampling and classifying as X or Y n_0+n_1 animals, the experimenter may tag or mark an equivalent number. The removal process will then act as a sample to estimate the tag ratio. It is perhaps reasonable to disregard the additional effort necessary to record accurately the tag recoveries. However, it is not reasonable to disregard the fact that tagging may be an operation that is more expensive, or requires more effort than classification.

In view of the result of the last section it is reasonable to consider the case $n_1 = n_2 = n$ (say). For simplicity we restrict consideration to $R_y = 0$, $R_x = R$. Denote R/N_0 by r. Then

$$\sigma^2(\hat{N}_0) = \frac{Y_0}{n}\left[\frac{X_0+X_0-R}{(1-P_0)^2}\right]\left(\frac{1-r}{r}\right)^2$$
$$= \frac{N_0^2}{n}\left[\frac{2P_0-r}{1-P_0}\right]\left[\frac{1-r}{r}\right]^2. \tag{11}$$

Now consider the estimate that could be made if $2\lambda n$ tags were placed in the population, where λ^{-1} is the ratio of the cost of tagging to the cost of classification and hence may be assumed to be greater than or equal to 1.

The random tag recoveries follow a hypergeometric distribution under the usual assumption of random sampling without replacement; this hypergeometric distribution is often approximated by a Poisson distribution (e.g. Chapman, 1948). Here, if n is considered to be small relative to N, while R is not necessarily so, then it may be shown by the usual procedures (using Stirling's formula for the large factorials appearing in the hypergeometric formula) that the distribution tends in the limit, as $N \to \infty$, while $n/N \to 0$ and R/N is bounded away from zero, to the binomial distribution with parameters $(n, R/N)$. If this limiting distribution is used as an approximation to the true hypergeometric distribution, the asymptotic variance of the estimate of N_0 based on the tag recoveries (say $\hat{N}_0(t)$) is

$$\sigma^2(\hat{N}_0(t)) = \frac{N_0^2(N_0-R)}{2\lambda nR} = \frac{N_0^2(1-r)}{2\lambda nr}. \tag{12}$$

Hence we may ask under what conditions on λ, r and P_0 is

$$\frac{1}{2\lambda} > \left(\frac{2P_0-r}{1-P_0}\right)\left(\frac{1-r}{r}\right). \tag{13}$$

This inequality holds provided

$$P_0 < \frac{r+2r\lambda(1-r)}{4\lambda(1-r)+r}. \tag{14}$$

Let

$$f(r,\lambda) = \frac{r+2r\lambda(1-r)}{4\lambda(1-r)+r}. \tag{15}$$

We observe that $f(r, 0) = 1$, while $f(r, 1)$ is a monotone function of r with $f(0, 1) = 0$, $f(1, 1) = 1$. The first statement in this sentence is the trivial remark that the change of composition method must always be better if tagging is prohibitively costly. The second case, where tagging is no more costly than classification, is of some interest; we consider this case in the next paragraph.

Elementary considerations show that for $r<1$, $P_0<f(r,1)$ implies $P_0<r$. Since the Y removal is zero, it is impossible for the proportion of the population removed to exceed the original proportion of X's. Consequently, if tagging is no more costly than classification, from the large sample point of view, the tag sample procedure is under all circumstances more efficient than the change of composition estimation method.

Turning to the case $\lambda>1$, we have used a different method of comparison. Tabled below is the minimum value of $1/\lambda$ such that $\sigma^2(\hat{N}_0(t))>\sigma^2(\hat{N}_0)$ for specified values of P_0.

Table 3. *Minimum value of relative cost of tagging to cost of classification such that* $\sigma^2(\hat{N}_0(t))>\sigma^2(\hat{N}_0)$

$X_0/N_0=P_0$	$R/N_0=r$						
	0·25	0·20	0·15	0·10	0·05	0·02	0·01
0·1	—	—	—	—	6·34	19·60	41·80
0·2	—	—	3·54	6·76	16·62	46·55	96·53
0·3	3·00	4·58	7·28	12·86	29·85	81·20	166·90
0·4	5·51	8·00	12·28	21·00	47·49	127·40	260·70
0·5	9·00	12·80	19·27	32·40	71·20	192·08	392·04
0·6	14·25	20·00	29·75	49·50	109·25	289·10	589·05
0·7	23·00	32·00	47·22	78·00	171·00	450·80	917·33
0·8	40·51	56·00	81·16	135·00	294·51	774·20	1574·10
0·9	93·00	128·00	87·00	306·00	332·51	1744·40	3544·20

Since the comparison of these two procedures is made under the conditions most favourable to the change of composition technique, these results suggest that in almost all cases the capture-recapture estimation procedure will yield more information for the same amount of effort. However, it is to be noted that the capture-recapture estimates are seriously dependent on the assumption that the initial capture does not alter the behaviour pattern of the animals in any way that will affect their probability of recapture. This assumption is difficult to test and it is certainly not always reasonable.

5. ESTIMATION WHEN R_x AND R_y MUST BE ESTIMATED

In the field, many additional complications may arise in the application of the change of composition estimation procedure. Often R is known reasonably well, but R_x and R_y are not; it may even happen that R itself must be estimated. These complications do not change the estimates but will modify the estimated variances and any interval estimates.

We study the case where R is known but R_x and R_y must be estimated from a sample of size m from the removed group. We assume that a sample of size m is taken randomly, and that of the m animals, m_x are from class X, m_y from Y. Then the estimates of X_0 and N_0, denoted by $\hat{X}_0$, $\hat{N}_0$, are simply (1) and (2) with R_x replaced by $\hat{R}_x$, where

$$\hat{R}_x=\frac{m_x R}{m}. \tag{16}$$

While the asymptotic variances of X_0 and N_0 may be determined by inverting a 3×3 information matrix, it is easier to compute them directly, for $\hat{X}_0$ and $\hat{N}_0$ are linear functions of R_x. Thus

$$\sigma^2(\hat{N}_0) = (P_0 - P_1)^{-2}\left[\left(\frac{X_0 Y_0}{n_0} + \frac{X_1 Y_1}{n_1}\right) + \frac{m R_x R_y}{R^2 (P_0 - P_1)^2}\left\{1 + 3\frac{P_0(1-P_0)}{n_0} + 3\frac{P_1(1-P_1)}{n_1}\right\}\right]. \quad (17)$$

This is the asymptotic variance, asymptotic in the sense that terms of higher order in n_0^{-1}, n_1^{-1} are neglected.

A similar formula can be derived for $\sigma^2(\hat{X}_0)$, and the same method can be followed through in the case that R itself has to be estimated in some manner. It is of course assumed that the estimation of R is independent of the random variables x_0, x_1. The assumption is implicit in the development above.

6. Combination of capture-recapture and change of composition estimates

It was noted in § 4 that, if the experimenter is prepared to accept the assumptions underlying the capture-recapture estimates, then he will gain more from putting his effort into tagging animals rather than merely classifying them. Actually, however, he may try to do both, for in this way he may get two different estimates, or the combined procedure may yield the most efficient single estimate. Moreover, the biologist will often wish to sample the population on more than one occasion, in order to study growth or other changes in the population. Consequently if the members of the initial sample are tagged as the basis of a tag-sample estimate, the normal pattern of the study of the population will make possible also the change of composition estimate (assuming of course a selective removal between the samples).

In this situation the first problem that confronts the experimenter is to determine whether there is reasonable agreement between the two possible estimates. As noted above in § 3, the binomial model with parameters $(n_0, R/N_0)$ is a reasonable approximation to represent the capture-recapture procedure. Since the large-sample variances are known for the change of composition estimate and for the estimate based on the number of tags recovered (cf. formula (12)), a large-sample test is available to compare the two separate estimates.

If the estimates are not compatible it may be due to failure in the assumptions in one or both models. These assumptions may be correct in regard to expectations, but the sources of variability may be greater than the models permit, i.e. the distributions are not binomial. This possibility will be considered in the next section.

On the other hand, there may be sources of error, such as were discussed in § 2, or are well known for capture-recapture estimates. These have been particularly well pointed out by De Lury (1954). It is pertinent to note that the most important sources of error tend to bias the change of composition estimate downward, while the important factors of trap-shyness and tag mortality will tend to inflate the tag sample estimates upwards.

If the estimates are compatible, there is a suggestion but certainly no proof that the assumptions of the two models are fulfilled. In this case it will be normal practice to combine them. They may be averaged, weighting them inversely to their asymptotic variances. However, for this simple model the maximum-likelihood equations are easily solved. We need to add to the notation of § 2 the following:

s = number of the n_0 animals taken in sample 0
which are recovered in the removal process.

It is assumed of course that these n_0 animals have been made identifiable by tagging or marking in some manner. Then

$$p(x_0, x_1, s) = \prod_{i=0}^{1} \binom{n_i}{x_i} \left(\frac{X_i}{N_i}\right)^{x_i} \left(\frac{Y_i}{N_i}\right)^{y_i} \binom{n_0}{s} \left(\frac{R}{N_0}\right)^{s} \left(1-\frac{R}{N_0}\right)^{n_0-s}. \tag{18}$$

The maximum-likelihood equations for X_0 and N_0 are seen to be of the same form as those of the simple model of § 2. They may be written in the form

$$\frac{x_0}{X_0} - \frac{2n_0}{N_0} + \frac{x_1}{X_1} - \frac{n_1 - n_0 + s}{N_1} = 0, \tag{19}$$

$$\frac{y_0}{Y_0} - \frac{2n_0}{N_0} + \frac{y_1}{Y_1} - \frac{n_1 - n_0 + s}{N_1} = 0, \tag{20}$$

and solved by analogy with those of the simple model. The maximum-likelihood estimators are

$$\hat{X}_0^c = \frac{x_0[(n_1 - n_0 + s)\, R_x - x_1 R]}{(n_1 - n_0 + s)\, x_0 - 2n_0 x_1}, \tag{21}$$

$$\hat{N}_0^c = \frac{2n_0[(n_1 - n_0 + s)\, R_x - x_1 R]}{(n_1 - n_0 + s)\, x_0 - 2n_0 x_1}. \tag{22}$$

The information matrix is

$$\begin{pmatrix} \dfrac{n_0}{X_0 Y_0} + \dfrac{n_1}{X_1 Y_1}, & -\dfrac{n_0}{N_0 Y_0} - \dfrac{n_1}{N_1 Y_1} \\ -\dfrac{n_0}{N_0 Y_0} - \dfrac{n_1}{N_1 Y_1}, & \dfrac{n_0 X_0}{N_0^2 Y_0} + \dfrac{n_1 X_1}{N_1^2 Y_1} + A \end{pmatrix}, \tag{23}$$

where

$$A = \frac{n_0 R}{N_0^2 N_1}. \tag{24}$$

Hence

$$\text{(Asymptotic) } \sigma^2(\hat{N}_0^c) = \left[\frac{(P_0 - P_1)^2}{\dfrac{X_0 Y_0}{n_0} + \dfrac{X_1 Y_1}{n_1}} + \frac{n_0 R}{N_0^2 N_1} \right]^{-1}. \tag{25}$$

An analysis of (25) shows that the optimum allocation of n_0 and n_1 is made by choosing n_0 the maximum possible, n_1 the minimum. This, of course, is necessarily true from the results obtained in § 4, if the differential cost of tagging is disregarded. Consequently the combined procedure outlined here will be used if there is reason to question the assumptions of the tag-sample procedure, or if further sampling after the removal process will be made for other reasons. In this case the question of optimum allocation does not enter.

7. EXTENSIONS FROM THE SIMPLE BINOMIAL MODEL

There are several immediate extensions or generalizations of the model considered in § 2. For example, there may be several samples with one removal or several removals with one pair of samples, or several samples and several removals. The first two of these extensions are trivial; the several selective removals between two random samples may simply be combined by addition and the original model is obtained. If there are several samples either before or after a single removal, they may be combined in the usual manner that samples pertaining to the binomial model are combined.

Where there are several selective removals, each followed by a random sample, a new situation arises. This is of some interest, for the sampling process might be combined with the removal process to lead to such a situation, i.e. the sampler might only return the members of the Y group to the population (though in general it is to be expected that the experimenter will prefer to return both X's and Y's marked or tagged).

The following additional notation is needed:

R_{x_i}: removal of X population prior to ith sample,

R_{y_i}: removal of Y population prior to ith sample,

$$R_i = R_{x_i} + R_{y_i} \quad (i = 0, 1, 2, \ldots, k).$$

Also $$X_i = X_0 - R_{x_i}, \quad N_i = N_0 - R_i, \quad \text{etc.}$$

The maximum-likelihood equations for X_0 and N_0 may be written down simply as

$$\left.\begin{aligned} \sum_{i=0}^{k} \frac{x_i}{X_i} &= \sum_{i=0}^{k} \frac{n_i}{N_i}, \\ \sum_{i=0}^{k} \frac{y_i}{Y_i} &= \sum_{i=0}^{k} \frac{n_i}{N_i}. \end{aligned}\right\} \tag{26}$$

These equations can be solved by iterative methods in any particular case. In case the R_{x_i}, R_{y_i} are small relative to X_0, Y_0, the following approximate solution may be employed to yield estimates without the labour of the iterative procedure. The logarithm of the likelihood function may be written

$$\begin{aligned} \ln L &= K + \sum_{i=0}^{k} x_i \ln\left(P_0 - \frac{R_{x_i}}{N_0}\right) + \sum_{i=0}^{k} y_i \ln\left(1 - P_0 - \frac{R_{y_i}}{N_0}\right) - \sum_{i=0}^{k} n_i \ln\left(1 - \frac{R_i}{N_0}\right) \\ &\doteq K + \sum_{i=0}^{k} x_i \ln P_0 - \sum_{i=0}^{k} \frac{x_i R_{x_i}}{N_0 P_0} + \sum_{i=0}^{k} y_i \ln(1 - P_0) - \frac{\Sigma y_i R_{y_i}}{(1 - P_0) N_0} + \sum_{i=0}^{k} \frac{n_i R_i}{N_0}, \end{aligned} \tag{27}$$

so that the likelihood equations become

$$(1 - \hat{P}_0) \sum_{i=0}^{k} x_i R_{x_i} + \hat{P}_0 \sum_{i=0}^{k} y_i R_{y_i} - \hat{P}_0 (1 - \hat{P}_0) \sum_{i=0}^{k} n_i R_i = 0, \tag{28}$$

$$\frac{\sum_{i=0}^{k} x_i}{\hat{P}_0} + \frac{\sum_{i=0}^{k} x_i R_{x_i}}{\hat{N}_0 \hat{P}_0^2} - \frac{\sum_{i=0}^{k} y_i}{1 - \hat{P}_0} - \frac{\sum_{i=0}^{k} y_i R_{y_i}}{(1 - \hat{P}_0)^2 \hat{N}_0} = 0. \tag{29}$$

$\hat{P}_0$ may be determined from the quadratic equation (28) and this in turn substituted in (29). The result is a simple linear equation for $\hat{N}_0$.

The possibility mentioned earlier that the selective removal might be performed by the sampler suggests a further possibility—to set up a sequential procedure for sampling, removing the X group in the sample from the population and so on for k steps where k is determined by the actual observations. This would be analogous to the sequential procedure proposed by Goodman (1953) for the tag sample method of estimation.

Perhaps more important than any of these extensions is to consider the situation where sampling that confirms to the binomial model is impossible. Schooling and segregation by size or sex or age will certainly be frequently encountered in many natural populations, so that factors beyond the control of the experimenter make invalid the assumptions of the

binomial model. This will be true even when he is cognizant of the principles of random sampling. Since these additional factors tend to produce more variability from sample to sample than is ascribable to the binomial model, it is necessary for the experimenter to replicate his samples to obtain an internal measure of this variability.

Consider then several samples $n_{ij}\,(i = 0, 1; j = 1, 2, \ldots, k_i)$, each composed of x_{ij} members of the X class and y_{ij} members of the Y class. To estimate X_0, N_0 a regression technique may be used that has some resemblance to De Lury's procedure for population estimates from the catch per unit of effort data (De Lury, 1947). To do so it is necessary to postulate that the random variables $p_{ij} = x_{ij}/n_{ij}$ are approximately normal with mean P_i and variance σ^2_{ij}.

Now define

$$\left.\begin{aligned} Z_{0j} &= X_0 - p_{0j}N_0, \\ Z_{1j} &= X_0 - R_x - p_{1j}(N_0 - R). \end{aligned}\right\} \tag{30}$$

Then $$\mathscr{E}(Z_{ij}) = 0, \quad \sigma^2(Z_{0j}) = N_0^2\sigma^2_{0j}, \quad \sigma^2(Z_{1j}) = (N_0 - R)^2\sigma^2_{1j}. \tag{31}$$

If the underlying heterogeneity is considerable it is reasonable to assume that $\sigma^2_{ij} = \sigma^2$ for all i, j; though if the n_{ij} differ considerably some note should be taken of this, presumably following the ideas of Cochran (1942). Also for simplification the factors N_0^2, $(N_0 - R)^2$ in the variances of the Z_{ij}'s are neglected. The maximum-likelihood equations for X_0 and N_0 then reduce to a form almost identical with (1) and (2), provided the single observations there are replaced by means.

Thus letting $p_{i.} = \dfrac{\sum_{j=1}^{k_i} p_{ij}}{k_i}$,

$$\hat{N}_0 = \frac{R_x - Rp_{1.}}{p_{0.} - p_{1.}}, \tag{32}$$

while by routine methods it is found that

$$\text{(Asymptotic) } \sigma^2(\hat{N}_0) = \frac{\sigma^2}{(P_0 - P_1)^2}\left[\frac{N_0^2}{k_0} + \frac{(N_0 - R)^2}{k_1}\right]. \tag{33}$$

It should be pointed out that (33) is a large sample variance if k_0, k_1 are large, not n_0, n_1 as was the case with respect to formulas (3) and (4).

Formula (33) may be used to estimate the magnitude of the sampling required for a given removal, or the magnitude of the removal for given sampling, to achieve an estimate with preassigned precision. Thus if $R_y = 0$, $k_0 = k_1 = k$ (say).

$$\text{(Asymptotic) } \sigma^2(\hat{N}_0) = \frac{1}{k}\frac{\sigma^2}{(1 - P_1)^2}\left[\frac{2 - 2r + r^2}{r^2}\right]. \tag{34}$$

Unfortunately, it will often be true in advance that so little is known as to the possible range of values for σ, that this formula cannot be used even to give qualitative results.

This procedure can be extended to the case of several selective removals, with a non-selective sample taken between each removal. No new theory is involved although the algebra becomes somewhat more tedious. However, such a process must necessarily take place over time where mortality and possibly emigration and immigration should be considered. It is therefore thought desirable to defer consideration of this case to a later study, where the restriction that the population is closed, is removed.

It is planned to illustrate the methods developed above in a further paper, using appropriate data concerned with fish populations.

REFERENCES

CHAPMAN, D. G. (1948). A mathematical study of confidence limits of salmon populations calculated from sample tag ratios. *Bull. Int. Pacif. Salm. Fish. Comm.* no. 2, pp. 67–85.

CHAPMAN, D. G. (1954). The estimation of biological populations. *Ann. Math. Statist.* **25**, 1–15.

COCHRAN, W. G. (1942). Sampling theory when the units are of unequal sizes. *J. Amer. Statist. Ass.* **37**, 199–212.

DE LURY, D. B. (1947). On the estimation of biological populations. *Biometrics*, **3**, 145–67.

DE LURY, D. B. (1954). On the assumptions underlying estimates of mobile populations. *Statistics and Mathematics in Biology*, pp. 287–93. Ames, Iowa: Iowa State College Press.

GOODMAN, LEO A. (1953). Sequential sampling tagging for population size problems. *Ann. Math. Statist.* **24**, 56–69.

SCATTERGOOD, L. W. (1954). Estimating fish and wildlife: a survey of methods. *Statistics and Mathematics in Biology*, pp. 273–85. Ames, Iowa: Iowa State College Press.

AN AGE-DEPENDENT BIRTH AND DEATH PROCESS

By W. A. O'N. WAUGH

Royal Military College of Science, Shrivenham

1. Introduction

Many examples of the stochastic processes known as branching or population processes have been developed. The assumptions leading to the various models have been decided partly by mathematical convenience and partly by the attempt to reproduce features controlling the development of biological or physical populations. The model we shall develop involves a continuous time parameter and an enumerable set of states, and we can compare it with several other processes of this type, e.g. the Yule–Furry birth process, Feller's birth and death process (see Feller, 1939, 1950), and the non-Markovian birth process due to D. G. Kendall (1948).

There are two motives for the development of the present model. Mathematically, it furnishes a conveniently soluble example of a non-Markovian process, in which transitions from a given state may be to one of several other states, and in this it is a generalization of Kendall's process in which transitions are always from the given state to its successor, i.e. from state number n to state number $n+1$. For purposes of application it is based on assumptions which are not proposed as ideal but which bring in features which are perhaps a step in the direction of realism. Thus it may help in the development of yet more realistic models when they are required, and in indicating both the scope and the limitations of the models constructed under an assumption which has been widely adopted in population problems; viz. that the individuals reproduce independently of one another.

Bellman & Harris (1948, 1952) have described a stochastic process similar to the one we shall examine, and we shall adopt their notation. They consider populations of particles which reproduce by fission independently of one another. They suppose that a 'generation time' is defined for these populations which is a random variable having a general distribution function $G(t)$ (which does not depend on the size of the population, in accordance with the assumption of independence just mentioned), so that $dG(t)$ expresses the probability that a particle born at time $t = 0$ ends its life in the interval $(t, t+dt)$.

They assume that with given probabilities $q_0, q_1, q_2, \ldots$ any particle whose life has just ended is replaced by $0, 1, 2, \ldots$ new particles.

Their detailed work is carried out for the pure birth process involving binary fission only:

$$q_2 = 1, \quad q_0 = q_1 = q_3 = \ldots = 0.$$

They state that their results can easily be generalized to other modes of fission. Note that when $q_0 \neq 0$ the process is a birth *and death* process. Further developments in this direction have also been obtained by Ramakrishnan (1951) and by Reid (1953).

It may be convenient to note here that a focal point of the present investigation will be the coefficient of variation of the population size, considered as a function of the time. The importance of this function in the study of populations has been emphasized previously, e.g. by Kendall (1952*a*).

2. Age-dependent process involving two possible fates: death or fission

For clarity and because it leads to a model which may be useful in applications, we shall not adopt, in the main part of our investigation, the most general assumptions that are possible. We shall indicate at various points throughout the paper certain further generalizations that merely involve difficulties of technique and alterations in detail which might be worked out if applications seemed to demand it.

We shall suppose that at the end of its life an individual either disappears or is replaced by two new individuals (i.e. death without issue or reproduction by binary fission). These two new individuals then develop and reproduce independently of one another in the same way as the parent individual, i.e. if the parent individual split at time t to form them, then the probability that either of the two new individuals ends its life during $(t+\tau, t+\tau+d\tau)$ is $dG(\tau)$, the distribution $G(\tau)$ being fixed, and the same for all individuals of the population. In the notation of Bellman & Harris this would be equivalent to putting

$$q_0 \neq 0, \quad q_2 \neq 0, \quad q_1 = q_3 = \ldots = 0,$$

but we shall make the more general assumption that these probabilities depend on the age at which the parent individual's life ends. A further generalization to admit the possibility of an individual's facing more than two alternative fates is quite simple. We shall suppose that at time $t = 0$ the population consists of just one *newly born* individual. Let us name the alternatives which face an individual of the population; risk (b) is the event that its life ends with the birth of two new individuals, risk (d) is the event that it dies without issue. We shall derive the functions $q_j(t)$ and $G(t)$ as follows. Suppose that there are two functions $\lambda(t)$ and $\nu(t)$ such that, given that an individual born at time 0 is alive at time t,

$$\Pr\{\text{individual succumbs to risk } (b) \text{ in } (t, t+dt)\} = \lambda(t)\,dt + o(dt),$$

$$\Pr\{\text{individual succumbs to risk } (d) \text{ in } (t, t+dt)\} = \nu(t)\,dt + o(dt),$$

thus

$$\Pr\{\text{individual's life ends in } (t, t+dt)\} = [\lambda(t)+\nu(t)]\,dt + o(dt).$$

The function $[\lambda(t)+\nu(t)]$ might be called the 'force of mortality'. In the argument which follows, and which leads to the fundamental integral equation, it is convenient if $\lambda(t)$ and $\nu(t)$ are both continuous and do not vanish together at any point in $0 < t < \infty$ and are such that $\int_0^t [\lambda(u)+\nu(u)]\,du \to \infty$ as $t \to \infty$. In the latter part of the paper we assume a certain form for $\lambda(t)$ and $\nu(t)$, and it happens that these requirements of continuity, etc., are fulfilled. However, it is possible to extend the argument to cover cases in which certain discontinuities are allowed, in particular the case where an instantaneous chain reaction can occur at the moment of fission (a possibility which is mentioned by Bellman & Harris (1952)). In this case our assumption of age-dependence permits the probabilities governing the numbers of progeny at instantaneous fission to be different to those corresponding to fission after a finite life length.

From $\lambda(t)$ and $\nu(t)$ we obtain the following conditional probabilities, given that the individual's life ends in $(t, t+dt)$:

$$\Pr\{\text{life ends by risk } (b)\} \equiv q_2(t) \equiv \frac{\lambda(t)}{\lambda(t)+\nu(t)},$$

$$\Pr\{\text{life ends by risk } (d)\} \equiv q_0(t) \equiv \frac{\nu(t)}{\lambda(t)+\nu(t)}.$$

Let $$G(t) \equiv 1 - \exp\left\{-\int_0^t [\lambda(u) + \nu(u)]\, du\right\}$$

$$= \Pr\{\text{individual born at time 0 lives for a time less than or equal to } t\}.$$

We can imagine an idealized population in which either the risk (d) is suspended and the population develops under risk (b) alone ($\nu(t) \equiv 0$) or vice versa. Then

$$B(t) \equiv 1 - \exp\left\{-\int_0^t \lambda(u)\, du\right\}$$

$$= \Pr\{\text{individual born at time 0 lives a shorter time than } t, \text{ when risk } (d) \text{ is suspended}\},$$

$$D(t) \equiv 1 - \exp\left\{-\int_0^t \nu(u)\, du\right\}$$

$$= \Pr\{\text{individual born at time 0 lives a shorter time than } t, \text{ when risk } (b) \text{ is suspended}\}.$$

If $$g(t) = \frac{dG(t)}{dt}, \quad d(t) = \frac{dD(t)}{dt}, \quad b(t) = \frac{dB(t)}{dt},$$

then $$q_0(t) = \frac{d(t)\,[1 - B(t)]}{g(t)}$$

and $$q_2(t) = \frac{b(t)\,[1 - D(t)]}{g(t)},$$

and also $$1 - G(t) = [1 - B(t)]\,[1 - D(t)].$$

3. The integral equation for the generating function of the population size

In order to obtain the mean and variance of the population size as functions of the time we start from an integral equation for the generating function of the distribution of the population size, which is very similar to the one from which Bellman & Harris (1952) derive their results. It is possible to give a definition of a suitable sample space Ω and a function $r_t(\omega)$ over Ω which may be called the population size at time t, and by considerations of measurability to show that $\{r_t\colon t \geqslant 0\}$ is a well-defined stochastic process, and furthermore to derive the fundamental integral equation from this definition. However, we will content ourselves with this reference to the more rigid approach and will merely give the following intuitive argument which leads to the integral equation.

Let $p_r(t)$ be the probability that the population size is r at time t, given that the population consisted of a single newly born individual at time 0.

Let $$\phi(z, t) \equiv \sum_{r=0}^{\infty} p_r(t)\, z^r \equiv E z^{r_t},$$

and let $$h(z, t) \equiv \sum_{n=0}^{\infty} q_n(t)\, z^n.$$

Then the situation at time t may have come about in either of two ways.

(i) The initial individual may still be alive.

(ii) The initial individual's life may have ended in the interval $(u, u + du)$ where $0 \leqslant u < t$. In this case, if its life ends in fate (b), then because of the independence of the members of the population and the fact that all individuals are assumed to develop under the same probabilities as regards life length and number of progeny, we have effectively two popula-

tions developing each for a period $t-u$, from initial ancestors newly born at time u. On the other hand, if its life ends in fate (d) then the population size at time t is zero.

Hence we have the integral equation

$$\phi(z,t) = \int_0^t [q_0(u) + q_2(u)\,\phi^2(z,t-u)]\,g(u)\,du + z[1-G(t)],$$

$$= \int_0^t h[\phi(z,t-u),u]\,g(u)\,du + z[1-G(t)],$$

the latter form of which can easily be seen to hold when modes of fission other than binary are considered.

In terms of the functions $B(t)$, $D(t)$ and their derivatives this can be written

$$\phi(z,t) = \int_0^t [\phi(z,t-u)]^2\,b(u)\,[1-D(u)]\,du + \int_0^t d(u)\,[1-B(u)]\,du + z[1-B(t)]\,[1-D(t)].$$

If we differentiate this equation with respect to z and then put $z=1$ we get the following equation for the mean population size at time t,

$$\mu_1(t) = 2\int_0^t \mu_1(t-u)\,b(u)\,[1-D(u)]\,du + [1-B(t)]\,[1-D(t)],$$

while if we differentiate twice with respect to z and then put $z=1$ we get the equation

$$\mu_2(t) = 2\int_0^t [\mu_1(t-u)]^2\,b(u)\,[1-D(u)]\,du + 2\int_0^t \mu_2(t-u)\,b(u)\,[1-D(u)]\,du,$$

for the second factorial moment of the population size.

4. The observable life-length distributions and the probability of extinction

In experimental observations we should probably be able to separate the data on life length into those for individuals ultimately suffering fate (b) and those for individuals ultimately suffering fate (d). These observable distributions will be related to the distributions $B(t)$ and $D(t)$ in a similar manner to that in which crude rates are related to net rates in a problem of competing risks.

Let $B^*(t)$ be the probability that an individual newly born at time 0 suffers fate (b) before time t. Then $B^*(\infty)$ is the probability that a newly born individual will ultimately suffer fate (b).

Let $B^*(t \mid b)$ be the probability that an individual newly born at time 0 and known ultimately to suffer fate (b) has a life length less than or equal to t.

Let $D^*(t)$, $D^*(\infty)$, $D^*(t \mid d)$ be the corresponding probabilities for fate (d).

The probability that an individual born at time 0 suffers fate (b) in the interval $(t, t+dt)$ is

$$[1-G(t)]\,\lambda(t)\,dt + o(dt).$$

Hence
$$B^*(t) = \int_0^t [1-G(u)]\,\lambda(u)\,du,$$

and, since
$$B^*(t) = B^*(\infty)\,B^*(t \mid b),$$

$$B^*(t \mid b) = \int_0^t [1-G(u)]\,\lambda(u)\,du \Big/ \int_0^\infty [1-G(u)]\,\lambda(u)\,du,$$

with a similar result for $D^*(t \mid d)$. It will be desirable in any attempt to fit the model to experimental data to give to $B^*(t \mid b)$ the form of the observed distribution of life lengths for those individuals whose life ultimately terminates in fate (b), and similarly for $D^*(t \mid d)$.

We can express $B^*(t)$ in terms of $B(t)$ and $D(t)$ as follows:

Since $$\lambda(t) = b(t)/[1-B(t)]$$

and $$1-G(t) = [1-B(t)][1-D(t)],$$

therefore $$B^*(t) = \int_0^t [1-D(u)]\, b(u)\, du,$$

and similarly $$D^*(t) = \int_0^t [1-B(u)]\, d(u)\, du.$$

From the rigorous definition of $Z(t)$ (the population size) as a random variable $r_t(\omega)$ over a suitable sample space Ω, to which we have referred, we can show that

$$p_0(t) \equiv \Pr\{Z(t)=0\}$$

is a solution of the integral equation

$$p_0(t) = \int_0^t \{q_0(u) + q_2(u)\,[p_0(t-u)]^2\}\, g(u)\, du.$$

Clearly $p_0(t)$ must be a bounded solution because we must have $0 \leqslant p_0(t) \leqslant 1$, and, because $\Pr\{Z(t_2)=0\} \geqslant \Pr\{Z(t_1)=0\}$ for $t_2 > t_1$, $p_0(t)$ is a monotonic increasing function of t. Furthermore, it can be shown, by a method due to Bellman & Harris, that the bounded solution of the integral equation is unique.

Hence $p_0(t) \uparrow P \leqslant 1$ as $t \to \infty$.

Furthermore, $p_0(t)$ is the sum of an indefinite integral and the convolution of a bounded monotonic increasing function with a continuous function. Hence $p_0(t)$ is continuous.

Let $p_0(t) = P - \delta(t)$ so that $\delta(t) \downarrow 0$ as $t \to \infty$.

The integral equation can be written

$$P - \delta(t) = \int_0^t \{q_0(u) + P^2 q_2(u)\}\, g(u)\, du - 2P \int_0^t \delta(t-u)\, q_2(u)\, g(u)\, du + \int_0^t \delta^2(t-u)\, q_2(u)\, g(u)\, du.$$

Note that $\int_0^t q_2(u)\, g(u)\, du \leqslant 1$, that $\delta(t) \leqslant 1$, and that by Cauchy's principle of convergence $\int_s^t q_2(u)\, g(u)\, du < \epsilon$, where ϵ is any positive constant, for any $t > s$ when s is sufficiently large. Consider the second term in the equation above and divide the range of integration into two parts

$$\begin{aligned}
&\int_0^s \delta(t-u)\, q_2(u)\, g(u)\, du + \int_s^t \delta(t-u)\, q_2(u)\, g(u)\, du \\
&\quad \leqslant \int_0^s \delta(t-u)\, q_2(u)\, g(u)\, du + \int_s^t q_2(u)\, g(u)\, du \quad \text{(because } 0 \leqslant \delta(t) \leqslant 1\text{)}, \\
&\quad \leqslant \int_0^s \delta(t-u)\, q_2(u)\, g(u)\, du + \epsilon \quad \text{for sufficiently large } s, \\
&\quad \leqslant \epsilon \int_0^s q_2(u)\, g(u)\, du + \epsilon \quad \text{for sufficiently large } t > s, \\
&\quad \leqslant 2\epsilon.
\end{aligned}$$

This argument and another similar one show that the last two terms on the right-hand side of the above integral equation tend to zero as $t \to \infty$, and so we get

$$P = \int_0^\infty \{q_0(u) + P^2 q_2(u)\}\, g(u)\, du.$$

Let $$\rho = \int_0^\infty q_0(u)\, g(u)\, du \quad \text{and} \quad \sigma = \int_0^\infty q_2(u)\, g(u)\, du,$$

then, since by the assumption that $\int_0^t [\lambda(u) + \nu(u)]\, du \to \infty$ as $t \to \infty$, the function $G(t) \to 1$ as $t \to \infty$

$$P = \rho + \sigma P^2, \quad \text{where} \quad \rho + \sigma = 1.$$

The roots of this quadratic are $P = 1$, $P = \rho/\sigma$. If $\rho \geqslant \sigma$ we must therefore have $p_0(t) \to 1$ because the limit of $p_0(t)$ is not greater than 1.

On the other hand, if $\rho < \sigma$, the limit of $p_0(t)$ might be either ρ/σ or 1.

Suppose the limit of $p_0(t)$ is 1. Then, for some t_1, $0 < t_1 < \infty$, we must have

$$p_0(t_1) = \theta, \quad \text{where} \quad \rho/\sigma < \theta < 1.$$

Then, for this value of t,

$$\theta = \int_0^{t_1} \{q_0(u) + q_2(u)\, [p_0(t_1 - u)]^2\}\, g(u)\, du$$

$$\leqslant \int_0^{t_1} \{q_0(u) + \theta^2 q_2(u)\}\, g(u)\, du.$$

Now if the upper limit of integration on the right-hand side is increased, the right-hand side cannot decrease. Furthermore, if $q_0(t)\, g(t)$ or $q_2(t)\, g(t)$ is non-zero for $t > t_1$, it must *increase*, if we increase the range of integration to $(0, \infty)$.

In this case

$$\theta < \rho + \sigma\theta^2,$$

i.e.

$$\sigma\theta^2 - \theta + \rho > 0.$$

But θ lies between the roots ρ/σ and 1 of the left-hand side, which is a quadratic expression that must be negative inside the interval $(\rho/\sigma, 1)$. Hence we have a contradiction, and therefore $p_0(t)$ cannot take any value in the interval $(\rho/\sigma, 1)$.

On the other hand, if $q_0(t)\, g(t)$ and $q_2(t)\, g(t)$ vanish identically for $t > t_1$, we may have the equality

$$\theta = \rho + \sigma\theta^2,$$

but this implies $\theta = \rho/\sigma$ or $\theta = 1$, which again contradicts the assumption $\rho/\sigma < \theta < 1$.

Hence we see that $p_0(t) \uparrow \min\{\rho/\sigma, 1\}$ and this limit is the probability of extinction.

5. Example: the simplest Markovian birth and death process

To illustrate the methods of the preceding sections we shall obtain the integral equation, etc., for the birth and death process in which the probabilities are as follows:

For a particle alive at time t

$$\text{probability to split into 2 new particles during } (t, t+dt) = \lambda\, dt + o(dt),$$

$$\text{probability to die during } (t, t+dt) = \nu\, dt + o(dt),$$

where λ and ν are constants.

In this case $b(t) = \lambda e^{-\lambda t}$ and $d(t) = \nu e^{-\nu t}$. Hence we obtain

$$g(t) = (\lambda+\nu)e^{-(\lambda+\nu)t},$$

$$q_0(t) = \nu/(\lambda+\nu), \quad q_2(t) = \lambda/(\lambda+\nu),$$

and the integral equation is

$$\phi(z,t) = \int_0^t [\nu + \lambda\phi^2(z, t-u)]\, e^{-(\lambda+\nu)u}\, du + z\, e^{-(\lambda+\nu)t}.$$

For the observable life-length distributions we have

$$\begin{aligned} B^*(t \mid b) &= \int_0^t \lambda\, e^{-(\lambda+\nu)u}\, du \Big/ \int_0^\infty \lambda\, e^{-(\lambda+\nu)u}\, du \\ &= 1 - e^{-(\lambda+\nu)t} \\ &= G(t), \end{aligned}$$

and the same result holds for $D^*(t \mid d)$. For the extinction probabilities since

$$\sigma = B^*(\infty) = \lambda/(\lambda+\nu) \quad \text{and} \quad \rho = D^*(\infty) = \nu/(\lambda+\nu),$$

the probability of extinction is equal to 1 if $\lambda \leqslant \nu$ and is equal to ν/λ otherwise. These particular results are of course all well known.

6. Choice of a particular form for $b(t)$ and $d(t)$

Although the functions that appear in the basic statement of the birth and death process are $\lambda(t)$ and $\nu(t)$, it is a clearer way of indicating the type of age dependence we are considering if we give first the frequency functions of the life-length distributions. In the ensuing sections of this paper we shall take

$$b(t) \equiv \frac{(k\lambda)^k}{(k-1)!}\, t^{k-1} e^{-k\lambda t},$$

$$d(t) \equiv \frac{(m\nu)^m}{(m-1)!}\, t^{m-1} e^{-m\nu t},$$

where λ and ν are positive constants.

Probability distributions of this type have been used before in population problems (see Kendall, 1952*b*) and also (much earlier) in the theory of queues (for an account of this see Erlang, 1948). A 'microscopic' interpretation of them has been given in the birth process, where only one fate is possible for each member of the population, i.e. in the binary fission process with no death effect. This interpretation is that at the birth of a new particle a process is started within it, which must go through a fixed number, k, of successive distinct phases. The times spent in the successive phases are independent random variables, each having distribution function $1 - e^{-k\lambda t}$. When the last phase is completed the particle subdivides and forms two new particles. It follows that the 'generation time' is distributed like $\chi^2_{2k}/(2k\lambda)$. Note that in giving this interpretation we do not intend to imply that if a real population is found to have a generation time with this distribution then a process of k phases is necessarily going on within the individuals. Our reason for adopting this form of distribution of generation time is that it appears to give a fairly good fit in certain populations (see, for example, Kendall (1948) and recent experimental results by Dr E. O. Powell (1955), and a real physical multiphase process may or may not be held to account for this.

We can extend this multi-phase interpretation to our model if we consider that at the birth of a new particle two such processes start within it, one being of k phases and the other of m phases. We suppose that the fate of the particle is decided as follows: if the k-phase process (which we might call 'maturity') is completed first, the particle undergoes fate (b) at the moment of completion, if the m-phase process (which we might call 'decay') is complated first, the particle undergoes fate (d).

An important property of the above distributions is that as the 'number of phases' k (or m) tends to infinity the distributions tend towards the corresponding deterministic distributions, while for all k (or m) they have the fixed mean value $1/\lambda$ (or $1/\nu$).

It follows from the above choice of $b(t)$ and $d(t)$ that

$$B(t) \equiv 1 - e^{-k\lambda t}\left[1 + k\lambda t + \ldots + \frac{(k\lambda t)^{k-1}}{(k-1)!}\right],$$

$$D(t) \equiv 1 - e^{-m\nu t}\left[1 + m\nu t + \ldots + \frac{(m\nu t)^{m-1}}{(m-1)!}\right],$$

$$\lambda(t) \equiv \frac{k\lambda}{(k-1)!}\frac{(k\lambda t)^{k-1}}{1 + k\lambda t + \ldots + \frac{(k\lambda t)^{k-1}}{(k-1)!}},$$

$$\nu(t) \equiv \frac{m\nu}{(m-1)!}\frac{(m\nu t)^{m-1}}{1 + m\nu t + \ldots + \frac{(m\nu t)^{m-1}}{(m-1)!}}.$$

Making use of properties of the above forms of $B(t)$ and $D(t)$ we shall express each integral equation in terms of convolutions of functions, and then reduce it to a linear differential equation with constant coefficients. The subsequent solution depends on solving the characteristic equation explicitly, and hence the equation for the mean population size.

In the next two sections we shall suppose that $k \geqslant 1$, but that $m = 1$. Roughly speaking this assumption means that births occur in an age-dependent manner but deaths occur at random. Because a complete solution is available and because the assumption may well be realistic in some applications, we give the explicit solution for the mean and derive the variance and coefficient of variation of the population size from it. We can compare our results (i) with Kendall's (1948) uniform multi-phase pure birth process by putting $\nu = 0$, and (ii) with Feller's Markovian birth and death process by putting $k = m = 1$. This leads, of course, to the simple form of this process in which the birth and death rates do not depend on the population size.

When the number of phases in both $B(t)$ and $D(t)$ exceeds 1 an explicit solution of the characteristic equation is not readily obtainable, so we give asymptotic formulae for the mean and variance of the population size in terms of the dominant root of the characteristic equation.

We shall find the following notation convenient.

Let
$$(\alpha) = \alpha e^{-\alpha t}.$$

Let
$$\int_0^t f(t-u)\alpha e^{-\alpha u}\,du = f(t) * (\alpha).$$

Thus $\alpha^2 t e^{-\alpha t} = (\alpha) * (\alpha)$. We shall write $(\alpha) * (\alpha) = (\alpha)^{*2}$, and, extending this analogy with powers,

$$\frac{\alpha^n}{(n-1)!} t^{n-1} e^{-\alpha t} = (\alpha) * (\alpha) * \ldots * (\alpha) = (\alpha)^{*n}.$$

Let
$$\int_0^t \frac{\alpha^n}{(n-1)!} u^{n-1} e^{-\alpha u} du = \int_0^t (\alpha)^{*n} du$$
$$= 1 * (\alpha)^{*n}$$

be the distribution function of $\frac{1}{2}\chi^2_{2n}/\alpha$, where 1 is the function that is constant and equal to unity for all t in $-\infty < t < \infty$.

Let $k\lambda + m\nu = \gamma$.

In this notation we can express the functions which occur in the integral equations as follows:

$$b(t)[1-D(t)] = \frac{(k\lambda)^k}{(k-1)!} e^{-\gamma t}\left[t^{k-1} + m\nu t^k + \ldots + \frac{(m\nu)^{m-1}}{(m-1)!} t^{m+k-2}\right]$$
$$= (k\lambda)^k\left[\frac{1}{\gamma^k}\frac{\gamma^k}{(k-1)!} t^{k-1} e^{-\gamma t} + m\nu \frac{k}{\gamma^{k+1}}\frac{\gamma^{k+1}}{k!} t^k e^{-\gamma t} + \ldots\right.$$
$$\left. + \frac{(m\nu)^{m-1}}{(m-1)!}\frac{k(k+1)\ldots(k+m-2)}{\gamma^{k+m-1}}\frac{\gamma^{k+m-1}}{(k+m-2)!} t^{k+m-2} e^{-\gamma t}\right]$$
$$= \left(\frac{k\lambda}{\gamma}\right)^k\left[(\gamma)^{*k} + \frac{m\nu}{\gamma}\frac{k}{1}(\gamma)^{*(k+1)} + \ldots\right.$$
$$\left. + \left(\frac{m\nu}{\gamma}\right)^{m-1}\frac{k(k+1)\ldots(k+m+2)}{(m-1)!}(\gamma)^{*(k+m-1)}\right].$$

Similarly
$$d(t)[1-B(t)] = \left(\frac{m\nu}{\gamma}\right)^m\left[(\gamma)^{*m} + \frac{k\lambda}{\gamma}\frac{m}{1}(\gamma)^{*(m+1)} + \ldots\right.$$
$$\left. + \left(\frac{k\lambda}{\gamma}\right)^{k-1}\frac{m(m+1)\ldots(m+k-2)}{(k-1)!}(\gamma)^{*(m+k-1)}\right].$$

The observable life-length distributions corresponding to probabilities of the type we have adopted are given by

$$B^*(t) = \left(\frac{k\lambda}{\gamma}\right)^k\left[1 * (\gamma)^{*k} + \frac{m\nu}{\gamma}\frac{k}{1} 1 * (\gamma)^{*(k+1)} + \ldots\right.$$
$$\left. + \left(\frac{m\nu}{\gamma}\right)^{m-1}\frac{k(k+1)\ldots(k+m+2)}{(m-1)!} 1 * (\gamma)^{*(k+m-1)}\right],$$

whence $B^*(\infty)$ follows on noting that $1 * (\gamma)^{*(k+j)}$ tends to 1 as $t \to \infty$ for $j = 0, 1, \ldots, m-1$. Hence we can see that the distribution function $B^*(t \mid b)$, and similarly $D^*(t \mid d)$, is the sum of constant multiples of several distribution functions each of χ^2 form, i.e. it is a mixture of χ^2 variables (see Robbins, 1948).

The probability of extinction is equal to 1 if

$$\left(\frac{k\lambda}{\gamma}\right)^k\left[1 + \frac{m\nu}{\gamma}\frac{k}{1} + \ldots + \left(\frac{m\nu}{\gamma}\right)^{m-1}\frac{k(k+1)\ldots(k+m-2)}{(m-1)!}\right] \leqslant \frac{1}{2}.$$

Note that when $k = m$ this is equivalent to $\nu \geqslant \lambda$.

The expression above consists of the first m terms of the negative binomial distribution whose generating function is $p^k(1-qs)^{-k}$ when we put $p \equiv k\lambda/\gamma$ and $q \equiv m\nu/\gamma$, and thus the given condition is equivalent to the statement that *the median of this negative binomial distribution shall be greater than or equal to m.*

7. The case of random death: mean population size

In this case, putting $m = 1$, we have

$$b(t)\,[1 - D(t)] = \left(\frac{k\lambda}{\gamma}\right)^k (\gamma)^{*k},$$

$$d(t)\,[1 - B(t)] = \frac{\nu}{\gamma}\left[(\gamma) + \frac{k\lambda}{\gamma}(\gamma)^{*2} + \ldots + \left(\frac{k\lambda}{\gamma}\right)^{k-1}(\gamma)^{*k}\right].$$

The corresponding observable life-length distributions are obtained as follows:

$$B^*(t) = \left(\frac{k\lambda}{\gamma}\right)^k 1 * (\gamma)^{*k}, \quad \text{whence} \quad B^*(\infty) = \left(\frac{k\lambda}{\gamma}\right)^k$$

and so

$$B^*(t \mid b) = 1 * (\gamma)^{*k},$$

which is of just the same form as $B(t)$ with the parameter $\gamma = k\lambda + \nu$ replacing $k\lambda$. The effect of deaths occurring at random, on the observable birth rate, is merely to alter the constant in the distribution without altering the form of the distribution whereby the dependence of the birth-rate on the age of the parent is expressed.

$$D^*(t) = \frac{\nu}{\gamma}\left[1 * (\gamma) + \frac{k\lambda}{\gamma} 1 * (\gamma)^{*2} + \ldots + \left(\frac{k\lambda}{\gamma}\right)^{k-1} 1 * (\gamma)^{*k}\right],$$

whence

$$D^*(\infty) = \frac{\nu}{\gamma}\left[1 + \frac{k\lambda}{\gamma} + \ldots + \left(\frac{k\lambda}{\gamma}\right)^{k-1}\right] = 1 - \left(\frac{k\lambda}{\gamma}\right)^k,$$

and so

$$D^*(t \mid d) = \frac{\nu}{\gamma}\left[1 - \left(\frac{k\lambda}{\gamma}\right)^k\right]^{-1}\left[1 * (\gamma) + \frac{k\lambda}{\gamma} 1 * (\gamma)^{*2} + \ldots + \left(\frac{k\lambda}{\gamma}\right)^{k-1} 1 * (\gamma)^{*k}\right].$$

Thus, as might be expected, although the actual death-rate is independent of the age of the individuals the effect of the age-dependent birth-rate is to produce an observable death-rate which *is* age-dependent.

The probability of extinction is equal to 1 if $\left(\frac{k\lambda}{\gamma}\right)^k \leqslant \frac{1}{2}$, i.e. if $\nu \geqslant k(2^{1/k} - 1)\lambda$, while if $\nu < k(2^{1/k} - 1)\lambda$, then

$$\text{Probability of extinction} = \left[1 - \left(\frac{k\lambda}{\gamma}\right)^k\right] \Big/ \left(\frac{k\lambda}{\gamma}\right)^k$$

$$= \left(1 + \frac{1}{k}\frac{\nu}{\lambda}\right)^k - 1.$$

In the case of random death the integral equation for the mean population size may be written in the 'convolution' notation as

$$\mu_1(t) = 2\left(\frac{k\lambda}{\gamma}\right)^k \mu_1(t) * (\gamma)^{*k} + 1 - \left(\frac{k\lambda}{\gamma}\right)^k 1 * (\gamma)^{*k} - \frac{\nu}{\gamma} 1 * \left[(\gamma) + \frac{k\lambda}{\gamma}(\gamma)^{*2} + \ldots + \left(\frac{k\lambda}{\gamma}\right)^{k-1}(\gamma)^{*k}\right],$$

where we have put $1 - G(t) = 1 - 1 * g(t)$.

Let us define a differential operator,

$$\delta \equiv 1 + \frac{1}{\gamma}\frac{d}{dt},$$

so that

$$\delta[f(t) * (\gamma)] \equiv f(t).$$

Put $t = 0$ in the integral equation and in the relations resulting from applying the operators $\delta, \delta^2, \ldots, \delta^{k-1}$ to it. This gives

$$\mu_1(0) = 1,$$

$$\delta\mu_1(0) = \frac{k\lambda}{\gamma},$$

.....

$$\delta^{k-1}\mu_1(0) = \left(\frac{k\lambda}{\gamma}\right)^{k-1},$$

whence it follows that

$$\left(\frac{d}{dt}\right)^j \mu_1(t)\Big|_{t=0} = (-\nu)^j \quad (j = 0, 1, \ldots, k-1).$$

While δ^k gives the differential equation of the mean

$$\delta^k\mu_1(t) = 2\left(\frac{k\lambda}{\gamma}\right)^k \mu_1(t) + 1 - \left(\frac{k\lambda}{\gamma}\right)^k + \frac{\nu}{\gamma}\left[1 + \frac{k\lambda}{\gamma} + \ldots + \left(\frac{k\lambda}{\gamma}\right)^{k-1}\right].$$

The constant terms sum to zero so we get

$$\left\{\left(1 + \frac{1}{\gamma}\frac{d}{dt}\right)^k - 2\left(\frac{k\lambda}{\gamma}\right)^k\right\}\mu_1(t) = 0.$$

The characteristic equation for the above differential equation is

$$\left(1 + \frac{x}{\gamma}\right)^k - 2\left(\frac{k\lambda}{\gamma}\right)^k = 0.$$

If ω is the primitive kth root of unity this gives

$$x = 2^{1/k}k\lambda\omega^j - \gamma, \quad \text{where} \quad j = 0, 1, \ldots, k-1.$$

These roots lie on a circle in the complex plane, centre $-\gamma$, radius $2^{1/k}k\lambda$. The root of largest real part is

$$x = 2^{1/k}k\lambda - \gamma = k(2^{1/k} - 1)\lambda - \nu.$$

To obtain an explicit solution for the mean population size it is convenient to consider the mean population size which results when there is no death effect ($\nu = 0$). In this case the system reduces to the uniform multi-phase birth process described by Kendall (1948, 1952*b*). If we call the mean population size $\mu_1^*(t)$ when $\nu = 0$ we see that it satisfies the equation

$$\left\{\left(1 + \frac{1}{k\lambda}\frac{d}{dt}\right)^k - 2\right\}\mu_1^*(t) = 0,$$

together with the end conditions

$$\mu_1^*(0) = 1 \quad \text{and} \quad \left(\frac{d}{dt}\right)^j \mu_1^*(0) = 0 \quad \text{for} \quad j = 1, 2, \ldots, k-1.$$

The solution of this system is

$$\mu_1^*(t) = \frac{1}{2k}\sum_{j=0}^{k-1} \frac{2^{1/k}\omega^j}{2^{1/k}\omega^j - 1} \exp\left[k(2^{1/k}\omega^j - 1)\lambda t\right]$$

$$= e^{-k\lambda t}\sum_{r=0}^{\infty} 2^r\left[\frac{(k\lambda t)^{rk}}{(rk)!} + \frac{(k\lambda t)^{rk+1}}{(rk+1)!} + \ldots + \frac{(k\lambda t)^{rk+k-1}}{(rk+k-1)!}\right].$$

It follows that if we put $\mu_1(t) = e^{-\nu t}\mu_1^*(t)$,

then $\mu_1(t)$ will satisfy the differential equation, together with the end-conditions, that we have derived for it above.

The value of the mean population size for large t depends on whether the dominant root of the characteristic equation, i.e. the real root $(2^{1/k}-1)k\lambda-\nu$ is less than, equal to, or greater than 0. We shall consider these cases in turn.

(i) $\nu > k(2^{1/k}-1)\lambda$. The mean population size tends to zero like $e^{-\alpha t}$, where α is a positive constant, and it follows from this or from the value of $B^*(\infty)$ that the probability of extinction is 1.

(ii) $\nu = k(2^{1/k}-1)\lambda$. The probability of extinction is again equal to 1 but we obtain a finite limiting mean population size

$$\mu_1(t) \to C_k, \quad \text{where} \quad C_k = \frac{2^{1/k}}{2k(2^{1/k}-1)}.$$

(iii) $\nu < k(2^{1/k}-1)\lambda$. The probability of extinction is equal to $\left(1+\frac{1}{k}\frac{\nu}{\lambda}\right)^k - 1 < 1$. The mean population size will increase exponentially and we can write the following asymptotic expression for it:

$$\mu_1(t) \sim \frac{2^{1/k}}{2k(2^{1/k}-1)} \exp\left[\{(2^{1/k}-1)k\lambda-\nu\}t\right].$$

8. The case of random death: coefficient of variation of the population size

Using the convolution notation we may express the integral equation connecting the mean with the second factorial moment of the population size as follows:

$$\mu_2(t) = 2\left(\frac{k\lambda}{\gamma}\right)^k \{\mu_1(t)\}^2 * (\gamma)^{*k} + 2\left(\frac{k\lambda}{\gamma}\right)^k \mu_2(t) * (\gamma)^{*k}.$$

By applying the differential operator δ as in the case of the equation for the mean this reduces to

$$\left\{\left(1+\frac{1}{\gamma}\frac{d}{dt}\right)^k - 2\left(\frac{k\lambda}{\gamma}\right)^k\right\}\mu_2(t) = 2\left(\frac{k\lambda}{\gamma}\right)^k \{\mu_1(t)\}^2,$$

i.e. multiplying through by γ^k and factorizing the operator

$$\prod_{r=0}^{k-1}\left\{\frac{d}{dt} - (2^{1/k}\omega^r - 1)k\lambda + \nu\right\}\mu_2(t) = 2(k\lambda)^k\{\mu_1(t)\}^2.$$

As with the mean, the behaviour of the second factorial moment as $t \to \infty$ depends on whether the dominant root of the characteristic equation is negative, zero, or positive, and we shall examine these cases in turn.

(i) $\nu > k(2^{1/k}-1)\lambda$. All terms in the solution for the second factorial moment will tend to zero faster than $e^{-\alpha t}$, where α is some positive constant.

(ii) $\nu = k(2^{1/k}-1)\lambda$. The term $\{\mu_1(t)\}^2$ on the right-hand side of the equation tends to the finite limit C_k^2 as $t \to \infty$, while the equation is

$$\frac{d}{dt}\prod_{r=1}^{k-1}\left\{\frac{d}{dt} - k(2^{1/k}\omega^r - 1)\lambda + \nu\right\}\mu_2(t) = 2(k\lambda)^k\{\mu_1(t)\}^2.$$

Hence for large t we shall have an asymptotic solution:

$$\mu_2(t) \sim \frac{2(k\lambda)^k}{\prod_{r=1}^{k-1}\{\nu - k(2^{1/k}\omega^r - 1)\lambda\}} C_k^2 t$$
$$= \frac{2(k\lambda)^k}{k(k\lambda+\nu)^{k-1}} C_k^2 t.$$

Hence the variance tends to infinity, but does so to the order of t rather than e^t and in this the behaviour of this age-dependent system resembles that of the simple Markovian birth and death process when the parameters λ and ν in it are equal.

Note that as $k \to \infty$ we have the condition $\nu = \lambda \log_e 2 \approx 0{\cdot}69315\lambda$ for a finite limiting mean population size.

(iii) $\nu < k(2^{1/k}-1)\lambda$. The dominant term in the complementary function will be of the order of $\exp[\{k(2^{1/k}-1)\lambda - \nu\}t]$, but the particular integral will contain a term in $\exp[\{k(2^{1/k}-1)\lambda - \nu\}2t]$ which will therefore give the asymptotic value of the complete solution. Formally we can write the particular integral as

$$\mu_2(t) = \frac{2(k\lambda)^k}{\left(\gamma + \frac{d}{dt}\right)^k - 2(k\lambda)^k}\{\mu_1(t)\}^2,$$

and hence

$$\mu_2(t) \sim \frac{2(k\lambda)^2}{[k(2^{(k+1)/k}-1)\lambda - \nu]^k - 2(k\lambda)^k}\{\mu_1(t)\}^2.$$

Let $V_n(t)$ be the coefficient of variation of the population size. Then

$$V_n(t) = \frac{\mu_2(t)}{\{\mu_1(t)\}^2} + \frac{1}{\mu_1(t)} - 1,$$

and if we introduce the asymptotic values for the mean and second factorial moment, we obtain for the coefficient of variation for large t

$$V_n^2 = \lim_{t\to\infty} V_n^2(t) = \frac{2}{\left(2^{(k+1)/k} - 1 - \frac{\nu}{k\lambda}\right)^k - 2} - 1$$
$$= \frac{2}{\left(1 + \frac{2\alpha_k - \nu/k}{k}\right)^k - 2} - 1,$$

where $\alpha_k \equiv k(2^{1/k}-1)$.

When $k \to \infty$, the process tends to resemble one in which deaths occur at random but fission occurs deterministically when the parent individual attains a certain fixed age equal to $1/\lambda$. The limit as $k \to \infty$ of the above coefficient of variation, which is itself a limit for large t, is thus of interest. Noting that $\alpha_k \to \log_e 2$ and that $\frac{1}{2} < e^{-\nu/\lambda} \leqslant 1$, with equality only when $\nu = 0$ we obtain the limit

$$\lim_{k\to\infty}(V_n^2) = 2\left(\frac{1 - e^{-\nu/\lambda}}{2e^{-\nu/\lambda} - 1}\right).$$

If V_τ is the coefficient of variation of the distribution $B(t)$, then $V_\tau = 1/\sqrt{k}$ and so $V_n/V_\tau = V_n\sqrt{k}$ will have a finite limit only if $V_n \to 0$, as $k \to \infty$, i.e. if $\nu = 0$. This point will be of significance if any attempt is made to estimate V_τ from observations of V_n, the latter being easier to observe, for instance, in bacterial populations.

9. Birth- and death-rates both age-dependent

We have seen that when $k>1$ and $m>1$ the instantaneous probabilities of death and of fission both depend on the age of the individual considered. We will first investigate the behaviour of the mean population size under these conditions. The integral equation for the mean population size is

$$\mu_1(t) = 2\left(\frac{k\lambda}{\gamma}\right)^k \mu_1(t) * \left[(\gamma)^{*k} + \frac{m\nu}{\gamma}\frac{k}{1}(\gamma)^{*(k+1)} + \ldots + \left(\frac{m\nu}{\gamma}\right)^{m-1}\frac{k(k+1)\ldots(k+m-2)}{(m-1)!}(\gamma)^{*(k+m-1)}\right] + [1-B(t)][1-D(t)].$$

Now
$$[1-B(t)][1-D(t)] \equiv e^{-\gamma t}P(t),$$
where $P(t)$ is a polynomial of degree $k+m-2$ in t. Hence

$$\delta^{k+m-1}[1-B(t)][1-D(t)] \equiv \left(1+\frac{1}{\gamma}\frac{d}{dt}\right)^{k+m-1} e^{-\gamma t}P(t) \equiv 0.$$

Thus, if we apply the operator δ^{k+m-1} to the above integral equation we obtain the following homogeneous differential equation for the mean population size

$$F\left(\frac{d}{dt}\right)\mu_1(t) \equiv \left\{\delta^{k+m-1} - 2\left(\frac{k\lambda}{\gamma}\right)^k\left[\delta^{m-1} + \frac{m\nu}{\gamma}\frac{k}{1}\delta^{m-2} + \ldots + \left(\frac{m\nu}{\gamma}\right)^{m-1}\frac{k(k+1)\ldots(k+m-2)}{(m-1)!}\right]\right\}\mu_1(t) = 0.$$

The characteristic equation is $F(x)=0$ and if we put $y = 1+x/\gamma$ this is

$$H(y) \equiv y^{k+m-1} - 2\left(\frac{k\lambda}{\gamma}\right)^k\left[y^{m-1} + \frac{m\nu}{\gamma}\frac{k}{1}y^{m-2} + \ldots + \left(\frac{m\nu}{\gamma}\right)^{m-1}\frac{k(k+1)\ldots(k+m-2)}{(m-1)!}\right] = 0.$$

It does not seem as easy in this case as it was in the case of random death ($m=1$) to obtain an explicit solution of the characteristic equation. We will obtain an asymptotic solution for the mean population size, and will consider only the case when the probability of extinction is less than 1. The condition for the probability of extinction to be less than 1 is $\rho<\sigma$ and this is the same as $1-2B^*(\infty)<0$, i.e. $H(1)<0$. Since $H(y)\to\infty$ as $y\to\infty$ it follows that the characteristic equation $H(y)=0$ must have at least one real root greater than 1. There is just one change of sign in $H(y)$ so by Descartes' rule of signs it follows that $H(y)=0$ has a unique real positive root, say w, which is simple and greater than 1.

We can show that w is larger than the real part of any other root of $H(y)=0$. Let $z \equiv u+iv$ be any complex root. We can ignore real roots and complex roots with $u \leqslant 0$. We have, rearranging $H(z)=0$

$$2\left(\frac{k\lambda}{\gamma}\right)^k\left\{\frac{1}{z^k} + \frac{m\nu}{\gamma}\frac{k}{1}\frac{1}{z^{k+1}} + \ldots + \left(\frac{m\nu}{\gamma}\right)^{m-1}\frac{k(k+1)\ldots(k+m-2)}{(m-1)!}\frac{1}{z^{k+m-1}}\right\} = 1,$$

whence, taking moduli,

$$2\left(\frac{k\lambda}{\gamma}\right)^k\left\{\frac{1}{|z|^k} + \frac{m\nu}{\gamma}\frac{k}{1}\frac{1}{|z|^{k+1}} + \ldots + \left(\frac{m\nu}{\gamma}\right)^{m-1}\frac{k(k+1)\ldots(k+m-2)}{(m-1)!}\frac{1}{|z|^{k+m-1}}\right\} \geqslant 1.$$

Now if z is complex, $v \neq 0$, and so $|z| > u$. The left-hand side, considered as a polynomial (in $|z|^{-1}$) is strictly increasing for any argument greater than 0, because all coefficients are positive, so if we substitute u^{-1} for $|z|^{-1}$ we can drop the sign of equality. Thus

$$2\left(\frac{k\lambda}{\gamma}\right)^k \left\{\frac{1}{u^k} + \frac{m\nu}{\gamma}\frac{k}{1}\frac{1}{u^{k+1}} + \ldots + \left(\frac{m\nu}{\gamma}\right)^{m-1} \frac{k(k+1)\ldots(k+m-2)}{(m-1)!}\frac{1}{u^{k+m-1}}\right\}$$

$$> 1 = 2\left(\frac{k\lambda}{\gamma}\right)^k \left\{\frac{1}{w^k} + \frac{m\nu}{\gamma}\frac{k}{1}\frac{1}{w^{k+1}} + \ldots + \left(\frac{m\nu}{\gamma}\right)^{m-1} \frac{k(k+1)\ldots(k+m-2)}{(m-1)!}\frac{1}{w^{k+m-1}}\right\},$$

whence $w > u$.

It follows that $F(x) = 0$ has a root r which is unique, simple, real and positive and larger than the real part of any other root of $F(x) = 0$. Hence we have, as the asymptotic expression for the mean population size

$$\mu_1(t) \sim c\, e^{rt},$$

where c is a constant which could in theory be evaluated, though to do so we should require the other roots of the characteristic equation, and the appropriate end-conditions. However, it is not essential to evaluate this constant, because it will disappear from the limiting form of the coefficient of variation of the population size.

To write down the differential equation for the second factorial moment of the population size we note that

$$\left(1+\frac{x}{\gamma}\right)^{k+m-1} - F(x) \equiv 2\left(\frac{k\lambda}{\gamma}\right)^k \left\{\left(1+\frac{x}{\gamma}\right)^{m-1} + \frac{m\nu}{\gamma}\frac{k}{1}\left(1+\frac{x}{\gamma}\right)^{m-2} + \ldots + \left(\frac{m\nu}{\gamma}\right)^{m-1}\frac{k(k+1)\ldots(k+m-2)}{(m-1)!}\right\}.$$

The integral equation for the second factorial moment is similar to the one which arose in the case of random death, but it contains more terms.

It can be reduced to a differential equation in just the same way, by applying the operator $\delta \equiv \left(1+\frac{1}{\gamma}\frac{d}{dt}\right)$, and the result is

$$F\left(\frac{d}{dt}\right)\mu_2(t) = \left\{\left(1+\frac{1}{\gamma}\frac{d}{dt}\right)^{k+m-1} - F\left(\frac{d}{dt}\right)\right\}\{\mu_1(t)\}^2.$$

We can see from the solution of the equation for the mean that the complementary function for this equation will have a dominant term of the order of e^{rt}, where r is the same dominant root of the characteristic equation. However, owing to the appearance of $\{\mu_1(t)\}^2$ on the right-hand side we see that the particular integral will contain a term of the order of e^{2rt} which will be the dominant term in the complete solution. Evaluating this term we obtain

$$\mu_2(t) \sim \frac{(1+2r/\gamma)^{k+m-1} - F(2r)}{F(2r)}\, c^2\, e^{2rt}.$$

Note that the denominator $F(2r) \neq 0$.

It follows that the coefficient of variation of the population size for large t is given by

$$V_n^2 = \lim_{t\to\infty} V_n^2(t) = \frac{(1+2r/\gamma)^{k+m-1}}{F(2r)} - 2.$$

In conclusion I should like to thank Mr D. G. Kendall for his continued help and encouragement throughout the preparation of this paper.

REFERENCES

BELLMAN, R. & HARRIS, T. E. (1948). On the theory of age-dependent stochastic branching processes. *Proc. Nat. Acad. Sci., Wash.*, **34**, 601–4.

BELLMAN, R. & HARRIS, T. E. (1952). On age-dependent binary branching processes. *Ann. Math.* **55**, 280–95.

ERLANG, A. K. (1948) *The Life and Works of A. K. Erlang.* Edited by E. Brockmeyer, H. L. Halstrøm and Arne Jensen. Copenhagen: *Trans. Danish Acad. Tech. Sciences*, 1948, No. 2.

FELLER, W. (1939). Die Grundlagen der Volterraschen Theorie des Kampfes ums Dasein in wahrscheinlichkeitstheoretischer Behandlung. *Acta biotheor., Leiden*, **5**, 11–40.

FELLER, W. (1950). *An Introduction to Probability Theory and its Applications.* New York: John Wiley and Sons, Inc.

HARRIS, T. E. (1951). Some mathematical models for branching processes. *Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability*, pp. 305–28.

KENDALL, D. G. (1948). On the role of variable generation time in the development of a stochastic birth process. *Biometrika*, **35**, 316–30.

KENDALL, D. G. (1952*a*). On the choice of a mathematical model to represent normal bacterial growth. *J. R. Statist. Soc.* B, **14**, 41–4.

KENDALL, D. G. (1952*b*). Les processus stochastiques de croissance en biologie. *Ann. Inst. Poincaré*, **13**, 43–108.

POWELL, E. O. (1955). Some features of the generation times of individual bacteria. *Biometrika*, **42**, 16–44.

RAMAKRISHNAN, ALLADI (1951). Some simple stochastic processes. *J. R. Statist. Soc.* B, **13**, 131–40.

REID, A. T. (1953). An age-dependent stochastic model of population growth. *Bull. Math. Biophys.* **15**, 361–5.

ROBBINS, H. (1948). Mixture of distributions. *Ann. Math. Statist.* **19**, 360–69.

A LARGE-SAMPLE BIOASSAY DESIGN WITH RANDOM DOSES AND UNCERTAIN CONCENTRATION*

By FRED C. ANDREWS and HERMAN CHERNOFF
The University of Nebraska and Stanford University

1. The Problem

The problem discussed here is that of designing an experiment for the purpose of estimating a parameter in a dose-response curve when the doses administered cannot be known with exactness.

To introduce the problem let us consider the following extreme example. A biologist has developed a strain of bacteria. He believes that this strain is so virulent that a dose of one organism applied to a test animal will lead to a response with a probability of the order of magnitude of 0·2. He wishes to estimate the virulence of this strain of bacteria more precisely. For experimentation he has available thirty test animals and 10 ml. of material containing this strain of bacteria in suspension. The concentration of bacterial organisms in this material is about four organisms per millilitre, but this latter number might be unreliable. The immediate problem confronting the biologist is one of design. Since he is not sure of the concentration he must use a portion of his material for a plate count (Clifton, 1950, pp. 243–4) to estimate the concentration. Then he must allocate portions of his remaining material among the test animals to determine virulence. The amount of material to be used in the plate count cannot be too large, for then the remainder will not be an adequate amount of dose material. Some balancing allocation between these two parts of the experiment must be reached or the entire experiment will not yield its maximum amount of information.

For the sake of the reader who prefers to peek at the last page of a mystery story, the optimal design corresponding to the biologist's problem (when completely formulated) is the following. To determine concentration 3·1 ml. of material are used. The remaining 6·9 ml. are divided equally among the thirty test animals.

According to this design about twenty-eight organisms are divided among thirty animals. It is obvious that not all of the animals are assured of receiving organisms. In fact, the dosage is random, and some animals might receive none, some might receive one and still others might receive two or more organisms. This points out that in our problem exact doses are not known for two reasons. First, the concentration is at best an estimate based on the limited amount of material assayed. Secondly, even if the concentration was known, the exact dose administered to an animal is a random variable which the biologist cannot observe directly. This second situation will be referred to as dosage subject to error. Dosage subject to error occurs frequently in practice but is seldom treated in theory (see Haley, 1953), although such theory is especially relevant when the doses are small.

The problem facing the biologist can be formulated briefly by stating that the fractions $f_0, f_1, f_2, \ldots, f_{30}$ must be determined with the understanding that f_0 is the fraction of the

* This work was sponsored by the U.S. Army, Navy and Air Force through the Joint Services Advisory Committee for Research Groups in Applied Mathematics and Statistics.

material to be used in the plate count and f_i the part to be used to dose the ith test animal. Of course these fractions are to be determined in such a manner that an 'optimal' design for estimating the dose response curve will result.

2. Model

In order to formulate completely this design problem it is necessary to specify carefully the nature of the probability distributions involved.

First, the dose-response curve must be specified. In this paper, we shall treat only the case where our biologist has reason to believe that if d organisms are administered to a test animal the probability of no response is given by

$$p_d = (1-\alpha)^d. \tag{1}$$

The parameter α represents the probability that a dose of one organism will lead to a positive response. This dose-response curve is said to be exponential because the dose d appears in the exponent. This formula may also be written

$$p_d = e^{d \log_e (1-\alpha)}. \tag{2}$$

This model has previously been discussed in the literature by Goldberg & Watkins (1952), Druett (1952) and Peto (1953), and in a different context by Cochran (1950).

Secondly, we must consider the distribution of the number of organisms appearing in a certain amount of the material. Frequently the law of small numbers is assumed in this situation (see Worcester, 1954) yielding: *If the available material is taken from a source where the concentration is c organisms per millilitre, the number of organisms x appearing in a sample of r millimetres is a Poisson random variable with mean cr. That is,*

$$P(x=i) = e^{-cr}\frac{(cr)^i}{i!} \quad (i = 0, 1, 2, \ldots). \tag{3}$$

Furthermore, if several samples are taken from the source (without replacement) the number of organisms in these samples are independently distributed.

As an example, suppose that one-half of the tube of 10 ml. were used in the plate count. Suppose also that the concentration of the source was exactly 4 organisms per millilitre. Then the number of organisms in the part plated would be a Poisson random variable with mean (expectation) $\frac{1}{2}.10.4 = 20$. When the dose d is a Poisson random variable with expectation D, we shall say that D is the nominal dose. Then

$$P(d=i) = \frac{e^{-D}D^i}{i!} \quad (i = 0, 1, 2, \ldots). \tag{4}$$

We might repeat here that in our problem the biologist does not know the concentration and must use the plate count to estimate the concentration of the source or equivalently the expected number of organisms in the test tube. Let λ be a notation for the expected number of organisms in the test-tube. If then the biologist uses a fraction f of his material as a dose, the nominal dose is $f\lambda$.

3. The effect of error in dose

As was pointed out in the introduction, the exact number of organisms in a dose is a random variable. Ordinarily, as in the case with the probit or logit models, the randomness of the dose very seriously complicates the mathematical situation, as is quite evident from reading

Haley (1953). It is, however, fortunate that with the use of the exponential model this complication is minimized.

Suppose that a nominal dose D is applied, then the probability of a negative response is given by

$$P_D = \sum_{i=0}^{\infty} P(d=i)\,(1-\alpha)^i = e^{-\alpha D}.$$

Contrast this with the probability of a negative response when the exact dose is d, namely,

$$p_d = (1-\alpha)^d = e^{[\log_e(1-\alpha)]d}.$$

It is fundamental here to see that the response curve for a nominal dose is also exponential in form. The only difference in the two curves is that $\log_e(1-\alpha)$ in the exact dose curve is replaced by $-\alpha$ in the nominal dose curve. (For small α, $\log_e(1-\alpha)$ is almost equal to $-\alpha$.)

4. THE NATURE OF AN ASYMPTOTICALLY LOCALLY OPTIMAL DESIGN. AN EXAMPLE

The problem we have posed in § 1 yields a solution which might be termed a large-sample locally optimal design. To illustrate what we mean by these terms, we shall use a simpler problem which has already been treated.

Suppose that it is desired to dose a large number of test animals with organisms from a material with *known concentration* and that there is an unlimited supply of organisms available. This problem reduces then to selecting nominal doses $D_1, D_2, D_3, \ldots$ to administer to the animals. Once such nominal doses are selected and observations obtained, one may apply maximum-likelihood or some other asymptotically efficient estimation technique. It is known (though not trivial) how to apply the maximum-likelihood method of estimation and how to obtain the asymptotic variance of the resulting estimate of α. By an asymptotically optimal design we mean one which minimizes the asymptotic variance. It turns out for the exponential model that the asymptotic variance is minimized when all the doses are the same and these are equal to

$$D = \frac{1{\cdot}6}{\alpha}.$$

Here the local character of the problem becomes evident. That is, the asymptotically optimal design to estimate α depends on α. But it is clear that if α were known, there would be no point to the experiment. This paradoxical situation is resolved by the following consideration. In practice, the biologist will have a rough idea of the order of magnitude of α. Suppose he felt he could guess α of the order of magnitude of 0·2, then he could use the dose 1·6/0·2 in place of $1{\cdot}6/\alpha$. It can easily be shown that the use of the wrong dose level increases the asymptotic variance of the estimate of α. However, this increase is very gradual. If we measure the dose in terms of percentage of the asymptotically optimal dose and the corresponding variance similarly we obtain the following table:

Dose/optimal dose	Variance/optimal variance
0·50	1·25
0·75	1·02
1·00	1·00
1·25	1·02
1·50	1·12

It is clear that if the biologist has any reasonable educated guess about α he could use the design which would be optimal if his guess were the correct value of α. If, however, his best guess might be off by a factor of three or more, alternative procedures would have to be considered. For example, he might use a small preliminary experiment involving several dose levels which are very highly spread out. Using the estimate of α based on this preliminary experiment he could then apply the locally optimal design.

The gradual increase in variance divided by optimal variance shows not only the applicability of the locally optimal design. It can also be used to show that designs where dose/optimal dose takes on several values from 0·75 to 1·25 are almost optimal. While such designs lead to more complexity in the calculation of maximum-likelihood estimates they are useful to biologists who have reservations about the applicability of the exponential model to their problem.

5. Results

As may have been expected, the large-sample optimal solution to our design problem involving an assay is a local solution. In fact, it will be shown in the appendix that the solution has the following characteristics:

(1) Every test animal is given an equal fraction of the test-tube. According to the dictates of the mathematical model there is no advantage in limiting the number of animals used to allow larger individual doses. Of course costs of obtaining individual animals would forestall the use of an unlimited number of animals in any one experiment.

(2) The fraction of the test-tube administered to test animals is given by β^0 which is a function of α (probability that one organism will cause a positive reaction), λ (the expected number of organisms in the test tube), and s (the number of test animals available).

This functional dependence is rather complicated, but β^0 is very well approximated by β^* which is obtained from the following simple formula:

$$\beta^0 \approx \beta^* = \text{minimum}\left(\frac{1}{1+\sqrt{\alpha}}, \frac{1{\cdot}6s}{\alpha\lambda}\right). \tag{5}$$

Hence if the numbers given in the statement were relevant we would have

$$\lambda \approx 40, \quad \alpha \approx 0{\cdot}2, \quad s = 30, \quad \frac{1}{1+\sqrt{\alpha}} \approx 0{\cdot}69, \quad \frac{1{\cdot}6s}{\alpha\lambda} \approx 6, \quad \beta^* \approx 0{\cdot}69.$$

A remark about how β^* was obtained might be pertinent here. If a very large number of organisms were available it would be desirable to give each animal a dose of about $1{\cdot}6/\alpha$ and to use the large amount remaining for the plate count. Then the fraction of the test-tube used for doses would be about $1{\cdot}6s/(\alpha\lambda)$. However, if there are very few organisms available, it turns out that $\sqrt{\alpha}/(1+\sqrt{\alpha})$ is approximately the right fraction to use in plate counting, leaving the fraction $1/(1+\sqrt{\alpha})$ for doses.

(3) The local nature of the solution is not a serious handicap to applications.

As in the example treated in § 3, the asymptotic variance undergoes a relatively small change even when the fraction devoted to doses is changed by a considerable amount. This can be seen by referring to Figs. 1 and 2 and Table 1. Table 1 gives β^0 as a function of α and λ for $s = 30$. In addition, there is presented $\underline{\beta}$, $\overline{\beta}$ and $\sigma^2(\hat{\alpha})$, where $\underline{\beta} < \beta^0 < \overline{\beta}$.

$\underline{\beta}$ and $\overline{\beta}$ are fractions devoted to doses which yield designs which are 80 % efficient (efficiency measured in terms of asymptotic variance). $\sigma^2(\hat{\alpha})$ is the asymptotic variance using β^0. The relatively large spread between $\underline{\beta}$ and $\overline{\beta}$ is an indication of the wide applicability of the local solution.

In Fig. 1, β, β^0, β^* and $\bar{\beta}$ are drawn as functions of λ for $\alpha = 0{\cdot}10$ and $s = 30$. Similarly, in Fig. 2, β, $\bar{\beta}^0$, β^* and $\bar{\beta}$ are obtained for $\alpha = 0{\cdot}04$. These figures indicate the goodness of the approximation of β^* to β^0.

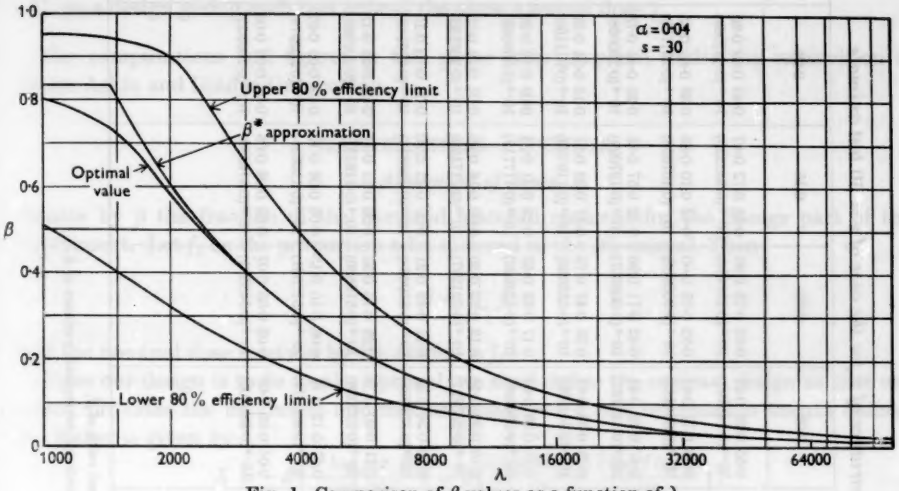


Fig. 1. Comparison of β values as a function of λ.

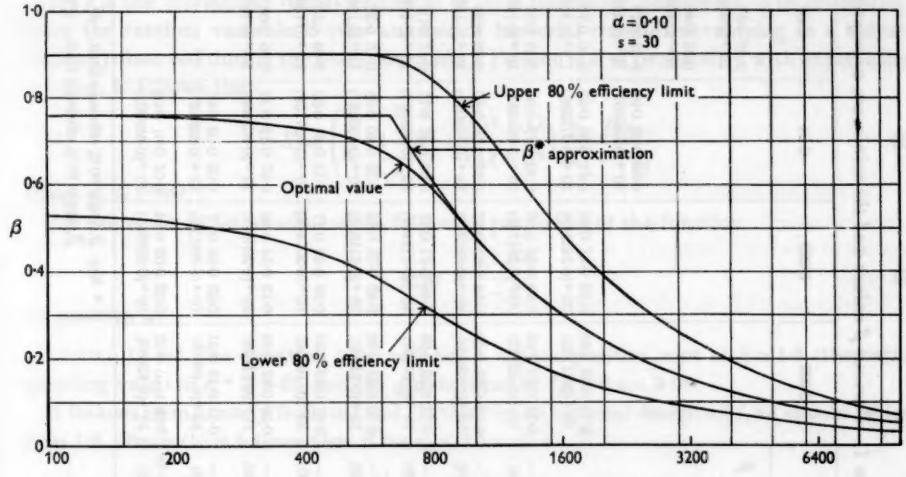


Fig. 2. Comparison of β values as a function of λ.

(4) The solution provides a very simple computational method of estimating α.

If all the test animals are given the same dose D, the proportion $\hat{p}$ of test animals which do not react is the maximum-likelihood estimate of $e^{-\alpha D}$. Hence, the estimate of α is given by

$$\hat{\alpha} = -\frac{1}{D}\log_e \hat{p} = \frac{-2{\cdot}30}{D}\log_{10}\hat{p}.$$

Table 1. *Various β values and the minimum asymptotic variance of $\hat{\alpha}$ for the case of $s = 30$ test animals*

λ \ α	0·001	0·005	0·020	0·04	0·06	0·08	0·10	0·15	0·20	λ
100	*$\underline{\beta}$ β^0 $\bar{\beta}$ $\sigma^2(\alpha)$				0·58 0·80 0·92 10^{-3}(0·993332)	0·55 0·78 0·91 10^{-2}(0·143101)	0·53 0·76 0·90 10^{-2}(0·191414)	0·48 0·72 0·87 10^{-2}(0·331126)	0·45 0·69 0·85 10^{-2}(0·497069)	100
250					0·56 0·80 0·92 10^{-3}(0·440869)	0·53 0·77 0·91 10^{-2}(0·654412)	0·50 0·75 0·90 10^{-2}(0·900903)	0·44 0·70 0·88 10^{-2}(0·167033)	0·38 0·65 0·85 10^{-2}(0·267765)	250
500				0·58 0·83 0·94 10^{-3}(0·147557)	0·52 0·79 0·93 10^{-3}(0·265671)	0·47 0·75 0·91 10^{-3}(0·417767)	0·42 0·71 0·90 10^{-2}(0·608201)	0·32 0·57 0·85 10^{-2}(0·127099)	0·25 0·45 0·73 10^{-2}(0·220868)	500
1,000		0·73 0·93 0·98 10^{-3}(0·61725)	0·62 0·87 0·96 10^{-4}(0·342318)	0·51 0·80 0·95 10^{-4}(0·973286)	0·40 0·70 0·92 10^{-3}(0·199053)	0·31 0·57 0·87 10^{-3}(0·344729)	0·25 0·47 0·75 10^{-3}(0·533669)	0·17 0·32 0·51 10^{-2}(0·119120)	0·13 0·24 0·39 10^{-2}(0·211150)	1,000
1,500	0·79 0·97 — 10^{-4}(0·7264)	0·73 0·93 0·98 10^{-3}(0·42732)	0·59 0·86 0·96 10^{-4}(0·264795)	0·40 0·72 0·94 10^{-4}(0·866463)	0·28 0·52 0·83 10^{-3}(0·190346)	0·21 0·39 0·64 10^{-3}(0·336499)	0·17 0·32 0·52 10^{-3}(0·524522)	0·12 0·21 0·34 10^{-2}(0·117719)	0·09 0·16 0·26 10^{-2}(0·209062)	1,500
2,000	0·79 0·97 — 10^{-4}(0·5492)	0·72 0·93 0·98 10^{-3}(0·33300)	0·54 0·85 0·97 10^{-4}(0·231948)	0·32 0·58 0·90 10^{-4}(0·843020)	0·22 0·39 0·64 10^{-3}(0·188297)	0·16 0·30 0·48 10^{-3}(0·333991)	0·13 0·24 0·39 10^{-3}(0·521297)	0·09 0·16 0·26 10^{-2}(0·117150)	0·07 0·12 0·19 10^{-2}(0·208160)	2,000
2,500	0·79 0·97 — 10^{-4}(0·4428)	0·71 0·93 0·98 10^{-3}(0·27695)	0·48 0·81 0·97 10^{-4}(0·217469)	0·26 0·47 0·77 10^{-4}(0·835757)	0·17 0·32 0·52 10^{-3}(0·187416)	0·13 0·24 0·39 10^{-3}(0·332788)	0·11 0·19 0·31 10^{-3}(0·519664)	0·07 0·13 0·20 10^{-2}(0·116871)	0·06 0·10 0·15 10^{-2}(0·207913)	2,500
3,000	0·78 0·97 — 10^{-4}(0·3720)	0·70 0·93 0·98 10^{-3}(0·24007)	0·42 0·75 0·96 10^{-4}(0·211656)	0·22 0·40 0·65 10^{-4}(0·832437)	0·15 0·26 0·43 10^{-3}(0·186968)	0·11 0·20 0·32 10^{-3}(0·332086)	0·09 0·16 0·26 10^{-3}(0·518686)	0·06 0·11 0·17 10^{-2}(0·116737)	0·05 0·08 0·13 10^{-2}(0·207336)	3,000
5,000	0·78 0·97 — 10^{-4}(0·2305)	0·67 0·93 0·98 10^{-3}(0·16956)	0·26 0·48 0·77 10^{-4}(0·207426)	0·13 0·24 0·39 10^{-4}(0·827759)	0·09 0·16 0·26 10^{-3}(0·186156)	0·07 0·12 0·19 10^{-3}(0·330874)	0·06 0·10 0·15 10^{-3}(0·517561)	0·04 0·06 0·10 10^{-2}(0·116535)	0·03 0·05 0·07 10^{-2}(0·206978)	5,000
10,000	0·77 0·97 — 10^{-4}(0·1250)	0·50 0·87 0·99 10^{-3}(0·13126)	0·13 0·24 0·39 10^{-4}(0·206413)	0·07 0·12 0·19 10^{-4}(0·825366)	0·05 0·08 0·13 10^{-3}(0·185690)	0·04 0·06 0·09 10^{-3}(0·330100)	0·03 0·05 0·07 10^{-3}(0·516391)	0·02 0·03 0·05 10^{-2}(0·116288)	0·02 0·02 0·04 10^{-2}(0·209933)	10,000
50,000	0·51 0·92 — 10^{-7}(0·518)	0·11 0·19 0·31 10^{-5}(0·12874)	0·03 0·05 0·07 10^{-4}(0·206220)	0·02 0·02 0·04 10^{-4}(0·838427)						50,000
100,000	0·26 0·48 0·78 10^{-7}(0·515)	0·06 0·10 0·15 10^{-5}(0·12886)	0·02 0·02 0·04 10^{-4}(0·209566)	0·01 0·01 0·02 10^{-4}(0·838262)						100,000

* Key: $\underline{\beta}$, lower 80% efficiency limit, β^0, optimal value of β,
$\bar{\beta}$, upper 80% efficiency limit, $\sigma^2(\alpha)$, minimum asymptotic variance of $\hat{\alpha}$.

Should a more complicated design be used, the calculation of the maximum-likelihood estimates would be much more difficult and involved. (It must be pointed out that if there are serious doubts about the applicability of the exponential model, the biologist cannot rely on a design giving each test animal the same nominal dose.)

The computations and figures in this paper were prepared under the supervision of Aloise Askin and Gladys Garabedian.

MATHEMATICAL APPENDIX

Allocation of doses

Denote by β the fraction of the bacterial material reserved for the dosage part of the experiment. Let f_i be the proportion administered to the ith animal. Then

$$\beta = \sum_{i=1}^{s} f_i$$

and the nominal dose received by ith animal is $f_i\lambda$.

Since our design is to be locally optimal, we shall define the optimal design as that one which provides the maximum information relative to α (the information matrix defined by Fisher is given by

$$X = -\mathscr{E}\left(\frac{\partial^2 \log f(x,\theta)}{\partial\theta_i \partial\theta_j}\right) = \mathscr{E}\left(\frac{\partial \log f(x,\theta)}{\partial\theta_i}\frac{\partial \log f(x,\theta)}{\partial\theta_j}\right),$$

where f is the probability density and $\theta = (\alpha, \lambda)$ is the vector parameter to be estimated). Since the random variable Z (the number of bacterial colonies developing in a culture medium) observed during the plate count has a Poisson law of probability with expectation $(1-\beta)\lambda$, it follows that

$$X = \begin{pmatrix} 0 & 0 \\ 0 & \dfrac{1-\beta}{\lambda} \end{pmatrix} + \begin{pmatrix} \lambda^2 & \alpha\lambda \\ \alpha\lambda & \alpha^2 \end{pmatrix} \sum_{i=1}^{s} \frac{f_i^2 p_i}{1-p_i}, \tag{6}$$

where $p_i = e^{-f_i\alpha\lambda}$.

For the moment let us assume the following properties of the function

$$g(u) = \frac{u^2 e^{-u}}{1-e^{-u}} \tag{7}$$

for positive u.

LEMMA 1. (*a*) The equation $g'(u) = 0$ has a unique positive root at $u \approx 1{\cdot}6$ (the corresponding value of e^{-u} is 0·8), and (*b*), $g(u)$ is concave for $0 \leqslant u \leqslant 3{\cdot}08$.

It follows immediately from (*a*) and (1) that for an optimal design all $f_i\alpha\lambda$ should be less than 1·6. From (*b*) it follows that if $0 \leqslant u_i \leqslant 3$ for all i

$$g\left(\frac{1}{s}\sum_{i=1}^{s} u_i\right) \geqslant \frac{1}{s}\sum_{i=1}^{s} g(u_i), \tag{8}$$

and therefore we have:

THEOREM 1. For the model considered here, all the animals should be given the same dose. It remains only to prove the lemma and to determine the optimal β.

Proof of Lemma. Part (*a*):

$$g'(u) = \frac{2u\,e^{-u}}{1-e^{-u}} - \frac{u^2 e^{-u}}{(1-e^{-u})^2}. \tag{9}$$

It suffices to show that the equation

$$(1-e^{-u})-\tfrac{1}{2}u=0, \tag{10}$$

has a unique positive root. But this follows from the fact that the expression on the left is concave and vanishes at $u=0$.

Part (*b*):

$$g''(u)=(e^u-1)^{-3}\left[e^{2u}(u^2-4u+2)+e^u(u^2+4u-4)+2\right].$$

Expanding e^{2u} and e^u in a Taylor series we get

$$g''(u)=(e^u-1)^{-3}\sum_{j=2}^{\infty}\frac{a_j u^j}{j!},$$

with $a_j=j(j-9)\,2^{j-2}+4(2^{j-2}-1)+j^2$, so that $a_j<0$ if $j=2,3,\ldots,8$ and $a_j>0$ if $j=9,10,\ldots$. From this it follows that $g''(u)$ has only one positive root (which is computed to be approximately 3·086) and (*b*) follows.

Suppose now that each animal is given the nominal dose $\beta\lambda/s$. The maximum-likelihood estimates of λ and α are

$$\hat{\lambda}=\frac{Z}{1-\beta} \tag{11}$$

and

$$\hat{\alpha}=\frac{-s}{\beta\hat{\lambda}}\log\hat{p}=\frac{-s}{Z}\frac{1-\beta}{\beta}\log\hat{p}, \tag{12}$$

where Z is the number of colonies developed during the plate count and $\hat{p}$ is the proportion of the animals which did not respond.

Studying $\hat{\alpha}$ directly or applying (6) it follows that the asymptotic variance of $\hat{\alpha}$ is

$$\sigma^2(\hat{\alpha})=\frac{\alpha^2}{\lambda(1-\beta)}+\frac{s}{\beta^2\lambda^2}\frac{(1-p)}{p},\quad p=e^{-\beta\lambda\alpha/s}. \tag{13}$$

The optimal value of β can be tabulated as a function of α and λ. Hence designs of this type can only be hoped to have local optimal properties. However, in many situations some *a priori* knowledge of λ, α is available which should enable one to get a fairly efficient design.

Examining $\sigma^2(\hat{\alpha})$ as a function of β we see that for small values of $\alpha\lambda/s$

$$\sigma^2(\hat{\alpha})\approx\frac{s}{\beta^2\lambda^2}\left[\frac{\beta\lambda}{s}\alpha\right]+\frac{\alpha^2}{\lambda(1-\beta)}=\frac{\alpha}{\lambda}\left\{\frac{1}{\beta}+\frac{\alpha}{1-\beta}\right\},$$

which is a minimum for $\beta=1/(1+\sqrt{\alpha})$.

For large values of λ there should be enough material to give each animal a dose of approximately ED_{80}, which is the optimal single dose as can be determined again from the information function. If this is the case, setting

$$\beta\lambda\alpha/s=1{\cdot}6$$

will yield an approximately optimal value for β when λ is large. For an overall approximation to optimal β we propose setting

$$\beta=\beta^*=\min\left[\frac{1}{1+\sqrt{\alpha}},\frac{(1{\cdot}6)s}{\alpha\lambda}\right].$$

Table 1 gives, for various values of α, λ, the actual value of β minimizing $\sigma^2(\hat{\alpha})$ for $s = 30$, the upper and lower limits of the values of β giving at least 80 % efficiency, and the minimum value of $\sigma^2(\hat{\alpha})$. If Table 1 is to be used for values of s other than $s = 30$ the conversion formula

$$\sigma^2(\hat{\alpha} \mid s, \lambda, \alpha, \beta) = \gamma\sigma^2(\hat{\alpha} \mid \gamma s, \gamma\lambda, \alpha, \beta) \tag{14}$$

is useful, where $\sigma^2(\hat{\alpha} \mid s, \lambda, \alpha, \beta)$ is the asymptotic variance of $\hat{\alpha}$ when s, λ, α, β are given. For example, if $s = 60$, $\lambda = 5000$, $\alpha = 0{\cdot}04$ and the minimum asymptotic variance and optimal β are desired, we see from Table 1 that for $s = 30$, $\lambda = 2500$, $\alpha = 0{\cdot}04$, the optimal β is $\beta = 0{\cdot}47$. From equation (14) we find that

$$\begin{aligned}\sigma^2(\hat{\alpha} \mid s = 60,\ \lambda = 5000,\ \alpha = 0{\cdot}4,\ \beta = 0{\cdot}47) \\ = \tfrac{1}{2}\sigma^2(\hat{\alpha} \mid s = 30,\ \lambda = 2500,\ \alpha = 0{\cdot}4,\ \beta = 0{\cdot}47) \\ = 10^{-4}(0{\cdot}835757).\end{aligned}$$

The optimal value of β is 0·47 and the upper and lower 80 % efficiency limits on β are 0·77 and 0·26, respectively.

If $\beta = \beta^*$ is used as an approximation to the optimal value of β, we see that a design of at least 80 % efficiency for all α, λ and an extremely efficient design for small λ and large λ result as can be seen in two instances in Figs. 1 and 2.

REFERENCES

CLIFTON, C. E. (1950). *Introduction to the Bacteria.* New York: McGraw Hill Book Co.

COCHRAN, W. G. (1950). Estimation of bacterial densities by means of the 'most probable number'. *Biometrics*, **6**, 105.

DRUETT, H. A. (1952). Bacterial invasion. *Nature, Lond.*, **170**, 288.

GOLDBERG, L. J. & WATKINS, H. M. S. (1952). A mathematical expression regarding mortality in the albino mouse to respiratory exposure with pathogenic micro-organisms. *Bact. Proc. Amer. Soc. Bact.* p. 74.

HALEY, K. D. C. (1953). Estimation of the dosage mortality relationship when the dose is subject to error. Unpublished Ph.D. dissertation, *Stanford University Library.*

PETO, S. (1953). A dose-response equation for the invasion of micro-organisms. *Biometrics*, **9**, 320.

WORCESTER, J. (1954). How many organisms? *Biometrics*, **10**, 227.

AN EXACT TEST FOR CORRELATION BETWEEN TIME SERIES

By E. J. HANNAN

Australian National University, Canberra, A.C.T.

1. Introduction

In a previous paper, Hannan (1955) in the process of obtaining an exact test for the serial correlation in the residuals from a regression, an estimator of the regression coefficient was also derived whose distributional properties, when the residual is a simple Gaussian Markoff process, are the same as those of the estimates of the regression coefficient in the classic cases. It was then shown that when the regressor is also generated by a simple Markoff process the regression coefficient will, for certain values of the serial correlation of the residual and the regressor, be a more efficient estimator than that obtained by straightforward least squares. It follows that in these cases the test of significance of the regression coefficient provides a test of the correlation between the two series which is asymptotically more powerful than the approximate test based on the sample correlation coefficient.

Alternative approximate tests for correlation between two series, in the form of partial correlations with the effects of lagged values of one or both variates removed, were proposed by Quenouille (1949). In the next section of this paper the efficiencies of Quenouille's tests are compared with those of the two above-mentioned tests when the residuals from the regression come from a simple Gaussian Markoff process.

In later sections the efficiencies of these tests are considered in other cases, some of which, from the point of view of a test for correlation, appear more interesting.

2. The exact test when the residuals from the regression come from a simple Markoff process independent of the regressor process

We will consider the regression

$$y_t = \alpha + \beta x_t + \epsilon_t \quad (t = 1, \ldots, n), \tag{1}$$

where (1) $\epsilon_t = \rho_1 \epsilon_{t-1} + \eta_t$ and η_t is $N\{0, \sigma_1(1-\rho_1^2)^{\frac{1}{2}}\}$, (2) ϵ_t is independent of x_s for all t and s. In a previous paper (Hannan, 1955) it was shown that an estimator of β could be obtained by considering the regression

$$y_{2t} = \alpha\left(1 - \frac{2\rho_1}{1+\rho_1^2}\right) + \frac{\rho_1}{1+\rho_1^2}(y_{2t-1} + y_{2t+1}) + \beta x_{2t} - \frac{\rho_1}{1+\rho_1^2}\beta(x_{2t-1} + x_{2t+1}) + \zeta_t \quad (t = 1, \ldots, [\tfrac{1}{2}(n-1)]).$$

Here ζ_t comes from a process of independent random variates with zero mean and variance $\sigma_1^2(1-\rho_1^2)/(1+\rho_1^2)$ and $[\frac{1}{2}(n-1)]$ is the greatest integer less than or equal to $\frac{1}{2}(n-1)$. The estimator, b_1, of β obtained from the sample regression coefficient of x_{2t} will have the usual properties of least-squares estimates on the normal case so that the corresponding

partial correlation between y_{2t} and x_{2t} will provide an exact test of the hypothesis that the two series are uncorrelated.

When x_t also comes from a simple Markoff process, with parameter ρ_2, ($|\rho_2|<1$), the variance of the estimator, b_1, was shown to be asymptotically

$$\frac{2\sigma_1^2(1-\rho_1^2)(1+\rho_2^2)}{n\sigma_2^2(1+\rho_1^2)(1-\rho_2^2)}.$$

Here σ_2^2 is the variance of x_t.

It is well known (Wold, 1953, p. 211) that the variance of the straightforward least-squares estimate, b_2, of β in the regression (1) is, asymptotically,

$$\frac{\sigma_1^2(1+\rho_1\rho_2)}{n\sigma_2^2(1-\rho_1\rho_2)}.$$

Alternative tests for correlation between two serially correlated series were proposed by Quenouille (1949). In the case of correlation between two Markoff processes he recommended the use of

$$r(x_2y_2\,|\,x_1), \quad r(x_2y_2\,|\,y_1) \quad \text{or} \quad r(x_2y_2\,|\,x_1y_1),$$

where $r(x_2y_2\,|\,x_1)$, for example, is the partial correlation between x_t and y_t with the effects of x_{t-1} removed. Asymptotically such statistics will have variances n^{-1}, when there is no correlation between the two series, since at least one of the series of residuals which are being correlated will approach independence. Quenouille showed, by a sampling experiment, that when the two series are uncorrelated the variances are approximately n^{-1}, even for small samples, and the bias is small. The best statistic, on these grounds, appeared to be $r(x_2y_2\,|\,x_1y_1)$. However, the efficiency of the statistics, as test statistics, depends not only on their variance on the null hypothesis but also on their distribution when the null hypothesis is not true. In particular, their expected values are relevant. A criterion of the asymptotic efficiency of a test based on a statistic t, of an hypothesis specified by a parameter value θ_0 is provided by the limit as $n\to\infty$ of the quantity (Stuart, 1954; Mood, 1954)

$$\left\{\frac{[\partial/\partial\theta\,\mathscr{E}(t)]^2}{\operatorname{var}(t)}\right\}_{\theta=\theta_0}. \tag{2}$$

The limit of the ratio of these quantities for two statistics t_1 and t_2 has been called (Pitman, 1948) the asymptotic relative efficiency of the tests and may be denoted by $E(t_1, t_2\,|\,\theta_0)$ (where the expression (2) for t_1 is in the numerator).

As mentioned, Quenouille was considering the correlation between two Markoff processes when he proposed his statistics. In the present case y_t will not be Markovian if $\rho_1\neq\rho_2$ and $\beta\neq 0$. It is, however, of interest to examine Quenouille's statistic in the present case also. (It is unlikely that, in practice, partial correlations of higher order than $r(x_2y_2\,|\,x_1y_1)$ will be used.)

We have

$$\lim_{n\to\infty}\left\{\frac{\partial}{\partial\beta}\mathscr{E}(r(x_2y_2\,|\,x_1))\right\}_{\beta=0} = \lim_{n\to\infty}\operatorname{cov}\{r(x_2y_2\,|\,x_1), \mathbf{y}'\boldsymbol{\Gamma}_1^{-1}\mathbf{x}\},$$

when β is zero (so that the y_t and x_t processes are independent). Here $\mathbf{y}$ and $\mathbf{x}$ are vectors of the n elements y_t and x_t and $\boldsymbol{\Gamma}_1$ is the covariance matrix of the residuals ϵ_t. The differentiation under the integral sign is justified by the uniform convergence of the resulting integral.

The limit of the covariance can be evaluated by straightforward methods (making use of an obvious extension of the theorem quoted in Cramer, 1946, p. 353) and gives

$$\lim_{n\to\infty}\left\{\frac{\partial}{\partial\beta}\mathscr{E}(r(x_2y_2\,|\,x_1))\right\}_{\beta=0}=\frac{\sigma_2(1-\rho_2^2)^{\frac{1}{2}}}{\sigma_1}.$$

Similarly

$$\lim_{n\to\infty}\left\{\frac{\partial}{\partial\beta}\mathscr{E}(r(x_2y_2\,|\,x_1y_1))\right\}_{\beta=0}=\frac{\sigma_2}{\sigma_1}\left(\frac{1-\rho_2^2}{1-\rho_1^2}\right)^{\frac{1}{2}}.$$

Since these two statistics have the same limiting variance on the null hypothesis $r(x_2y_2\,|\,x_1y_1)$ is always asymptotically the more efficient.

We now have

$$E(b_1,b_2\,|\,\beta=0)=\frac{(1+\rho_1\rho_2)(1+\rho_1^2)(1-\rho_2^2)}{2(1-\rho_1\rho_2)(1-\rho_1^2)(1+\rho_2^2)},$$

$$E\{r(x_2y_2\,|\,x_1y_1),b_2\,|\,\beta=0\}=\frac{(1+\rho_1\rho_2)(1-\rho_2^2)}{(1-\rho_1\rho_2)(1-\rho_1^2)},$$

$$E\{r(x_2y_2\,|\,x_1y_1),b_1\,|\,\beta=0\}=2\left(\frac{1+\rho_2^2}{1+\rho_1^2}\right).$$

The first of these ratios was shown for certain values of ρ_1 and ρ_2 in Table 3 in Hannan (1955).* The second ratio is shown for the same values of ρ_1 and ρ_2 in Table 1.

Table 1. $E(r(x_2y_2\,|\,x_1y_1),b_2\,|\,\beta=0)$

ρ_2 \ ρ_1	−0·8	−0·6	−0·4	−0·2	0	0·2	0·4	0·6	0·8
0	2·78	1·56	1·19	1·04	1	1·04	1·19	1·56	2·78
0·2	1·93	1·18	0·97	0·92	0·96	1·08	1·34	1·91	3·68
0·4	1·20	0·80	0·72	0·75	0·84	1·03	1·38	2·14	4·53
0·6	0·62	0·47	0·47	0·52	0·64	0·85	1·24	2·13	5·06
0·8	0·22	0·20	0·22	0·27	0·36	0·52	0·83	1·60	4·56

It is evident that, while b_1 will be asymptotically more powerful than b_2 when ρ_1 is sufficiently high it is always asymptotically less efficient than $r(x_2y_2\,|\,x_1y_1)$. It has the advantage, however, of giving an exact test.

It is easily shown that when x_t is Markovian $-\mathscr{E}\left(\frac{\partial^2\log L}{\partial\beta^2}\right)_{\beta=0}$ tends to

$$\{n\sigma_2^2(1+\rho_1^2-2\rho_1\rho_2)\}/\{\sigma_1^2(1-\rho_1^2)\}$$

as n increases; while the quantities $-\mathscr{E}\left(\frac{\partial^2\log L}{\partial\beta\,\partial\sigma_1^2}\right)_{\beta=0}$ and $-\mathscr{E}\left(\frac{\partial^2\log L}{\partial\beta\,\partial\rho_1}\right)_{\beta=0}$ tend to zero. Here L is the likelihood function. It follows (Whittle, 1953) that the variance of the maximum-likelihood estimator of β is, asymptotically, when $\beta=0$

$$\frac{\sigma_1^2(1-\rho_1^2)}{n\sigma_2^2(1+\rho_1^2-2\rho_1\rho_2)}.$$

None of the statistics considered in this section are, therefore, asymptotically fully efficient in general, though $r(x_2y_2\,|\,x_1y_1)$ is asymptotically fully efficient when $\rho_1=\rho_2$.

* There ρ and ρ_1 were used in place of ρ_1 and ρ_2 respectively.

One can, of course, suggest tests which are asymptotically fully efficient and which do not require the solution of non-linear systems of equations as the maximum-likelihood estimator does. Such a test will be obtained by using a consistent estimator of the serial correlation of the residuals to transform the regression equation to a form with an approximately random residual (Cochrane & Orcutt, 1949). Apart from the fact that the criterion of asymptotic relative efficiency is only suggestive of the relative power of tests in small samples it has an added deficiency when applied to tests which are not exact, for it does not take into account the effect of the deviation of the significance point used from the true value appropriate to the level of significance required. In the case of Quenouille's tests his sampling experiments seem to show that on the null hypothesis the mean and variance of his statistics in small samples are near to their theoretical values so that the significance point used will be approximately correct.

Example 1. Two series of seventy observations were chosen from series 7 in Kendall (1949), the last term in one series being sufficiently far from the first in the next to make the two series effectively independent. If the two series are represented by x_t and z_t, a third series which was formed can be represented by

$$y_t = 0{\cdot}8x_t + z_t.$$

The commencing term in the series x_t is no. 191 and in z_t no. 316.

Table 2

Statistic	Sample value	Population value	Test statistic
b_1	0·994	0·8	$t_{30} = 5{\cdot}27^{**}$
b_2	0·417	0·8	$r_{30} = 0{\cdot}316$
$r(x_2y_2 \mid x_1)$	0·316	0·329	$r_{67} = 0{\cdot}361^{**}$
$r(x_2y_2 \mid x_1y_1)$	0·684	0·625	$r_{66} = 0{\cdot}684^{**}$

** Highly significant.

The three series now are Gaussian Markoff processes with first serial correlation equalling 0·9.

The details of the various tests are given in Table 2.

Here t_{30} indicates Student's t with 30 degrees of freedom, while r_m indicates a correlation coefficient with m degrees of freedom.

The observed first serial correlation coefficients were

$$r_{1,x} = 0{\cdot}755, \quad r_{1,y} = 0{\cdot}739.$$

The degrees of freedom appropriate to r_{20} were calculated from the formula

$$n' = 70\frac{1-(0{\cdot}755)(0{\cdot}739)}{1+(0{\cdot}755)(0{\cdot}739)} \approx 20.$$

The statistic r_{20} is not significant at the 5 % point.

3. An exact test for correlation between two simple Markoff processes

Let
$$(y_t-\mu_1)=\rho_1(y_{t-1}-\mu_1)+\epsilon_t,$$
$$(x_t-\mu_2)=\rho_2(x_{t-1}-\mu_2)+\eta_t,$$

be two simple Markoff processes with normally distributed disturbances having zero means and variances $\sigma_1^2(1-\rho_1^2)$ and $\sigma_2^2(1-\rho_2^2)$.

The correlation between these processes can be prescribed in terms of the correlation properties of the residuals. Since these two residuals come from processes of independent random variates this will be achieved by saying for what lag, if any, the two residuals are correlated and what these correlations are.

We will consider the case where the only non-zero correlation is between ϵ_t and η_{t-m} and the series have been so lagged relative to each other that m is zero.

If the correlation between ϵ_t and η_t is ρ, then the correlation between x_t and y_s is

$$t\geqslant s: \quad \frac{\rho\{(1-\rho_1^2)(1-\rho_2^2)\}^{\frac{1}{2}}}{(1-\rho_1\rho_2)}\rho_2^{t-s}=\alpha\rho_2^{t-s},$$
$$t\leqslant s: \quad =\alpha\rho_1^{s-t}.$$

The two series y_t and x_t will be jointly normally distributed for every t so that

$$(y_t-\mu_1)=\alpha\frac{\sigma_1}{\sigma_2}(x_t-\mu_2)+\zeta_t,$$

where ζ_t is independent of x_t and has mean zero and variance $\sigma_1^2(1-\alpha^2)$.

Then a simple calculation shows that $\mathscr{E}(\zeta_t\zeta_{t-s})=\sigma_1^2(1-\alpha^2)\rho_1^s$. It follows, since ζ_t is normally distributed, that it is generated by a simple Gaussian Markoff process with parameter ρ_1 (Doob, 1953, p. 233). The disturbance θ_t in the process $\zeta_t=\rho_1\zeta_{t-1}+\theta_t$ is

$$\theta_t=\alpha\frac{\sigma_1}{\sigma_2}(\rho_1-\rho_2)(x_{t-1}-\mu_2)+\epsilon_t-\alpha\frac{\sigma_1}{\sigma_2}\eta_t.$$

If $\rho_1\neq\rho_2$ and $\rho\neq 0$, θ_t depends upon x_{t-1} and ζ_t is not independent of x_s for $t\neq s$. The θ_t are of course independent.

When $\rho_1=\rho_2$ the present case is equivalent to that considered in the previous section. When $\rho_1\neq\rho_2$ the exact test of significance of the correlation between the two series there given is still an exact test, but the conclusions as to the power of this test and the others there considered, when x_t also comes from a Markoff process, do not follow, for when $\rho\neq 0$ the condition (2) upon which the derivation depended does not hold.

The conditional distribution of the y_{2t} ($t=1,\ldots,[\frac{1}{2}(n-1)]$), for fixed x_t ($t=1,\ldots,n$) and y_{2t+1} ($t=0,\ldots,[\frac{1}{2}(n-1)]$) is the product of the conditional distributions of each y_{2t} for fixed y_{2t-1}, y_{2t+1}, x_{2t-1}, x_{2t} and x_{2t+1}. This follows from the fact that for fixed y_{2t+1} and x_{2t+1} the y_{2t} and x_{2t} are serially independent (Ogawara, 1951). This conditional distribution can be found

from the correlation matrix of the variates $y_{2t-1}, y_{2t+1}, x_{2t-1}, x_{2t}, x_{2t+1}, y_{2t}$ (taken in that order), which is

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & 1 \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} 1 & \rho_1^2 & \alpha & \alpha\rho_2 & \alpha\rho_2^2 \\ \rho_1^2 & 1 & \alpha\rho_1^2 & \alpha\rho_1 & \alpha \\ \alpha & \alpha\rho_1^2 & 1 & \rho_2 & \rho_2^2 \\ \alpha\rho_2 & \alpha\rho_1 & \rho_2 & 1 & \rho_2 \\ \alpha\rho_2^2 & \alpha & \rho_2^2 & \rho_2 & 1 \end{bmatrix},$$

$$\mathbf{B}' = \{\rho_1 \quad \rho_1 \quad \alpha\rho_1 \quad \alpha \quad \alpha\rho_2\}.$$

Then the variate y_{2t} is conditionally distributed with mean $\sigma_1 \mathbf{A}^{-1}\mathbf{B}Z$ and variance $\sigma_1^2\{1 - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}\}$ (Rao, 1952, p. 54).

Here
$$\mathbf{Z}' = \left\{\frac{y_{2t-1}}{\sigma_1} \quad \frac{y_{2t+1}}{\sigma_1} \quad \frac{x_{2t-1}}{\sigma_2} \quad \frac{x_{2t}}{\sigma_2} \quad \frac{x_{2t+1}}{\sigma_2}\right\}.$$

The vector $(\mathbf{A}^{-1}\mathbf{B})'$ is easily seen to be

$$\frac{1}{1+\rho_1^2}\left\{\rho_1 \quad \rho_1 \quad \frac{-\rho_2\alpha(1-\rho_1\rho_2)}{1-\rho_2^2} \quad \frac{\alpha(1-\rho_1^2\rho_2^2)}{1-\rho_2^2} \quad \frac{-\rho_1\alpha(1-\rho_1\rho_2)}{1-\rho_2^2}\right\},$$

while
$$\sigma_1^2\{1 - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}\} = \frac{\sigma_1^2(1-\rho^2)(1-\rho_1^2)}{1+\rho_1^2}.$$

For $\rho_1 = \rho_2$ these formulae are equivalent to those given in the previous section, when it is remembered that there the variance of the residual ϵ_t was σ_1^2, while here the variance of the corresponding residual is $\sigma_1^2(1-\rho^2)$.

For $\rho_1 \neq \rho_2$ two things are noticeable:

(1) The regression coefficient of x_{2t-1} is not equal to that for x_{2t+1} so that some of the symmetry present when $\rho_1 = \rho_2$ is now lost. This loss of symmetry is also evident from the fact that

$$\mathscr{E}\{(x_t - \mu_2)(y_{t+s} - \mu_1)\} \neq \mathscr{E}\{(x_{t+s} - \mu_2)(y_t - \mu_1)\},$$

when
$$\rho_1 \neq \rho_2 \quad \text{and} \quad s \neq 0$$

and
$$\mathscr{E}\{(y_t - \mu_1)(x_{t+s} - \mu_2) \mid y_{t+1}, y_{t-1}\} = 0 \quad s < 0,$$
$$\neq 0 \quad s > 0.$$

(2) The coefficient of x_{2t} is no longer equal to $\alpha\sigma_1/\sigma_2$ but instead to

$$\alpha\frac{\sigma_1}{\sigma_2}\left\{\frac{1-\rho_1^2\rho_2^2}{(1+\rho_1^2)(1-\rho_2^2)}\right\}.$$

The asymptotic variance of the estimator, b_1, of this last coefficient is, for $\rho = 0$,

$$\frac{2\sigma_1^2(1-\rho_1^2)(1+\rho_2^2)}{n\sigma_2^2(1+\rho_1^2)(1-\rho_2^2)},$$

so that the quantity
$$\lim_{n\to\infty}\left\{\operatorname{var}(b_1)\Big/\left[\frac{\partial}{\partial\rho}\mathscr{E}(b_1)\right]^2\right\}_{\rho=0}$$

is
$$\frac{2(1+\rho_1^2)(1+\rho_2^2)}{n(1+\rho_1\rho_2)^2}.$$

For the ordinary correlation coefficient, r, between x_t and y_t the corresponding quantity is (Bartlett, 1935)

$$\frac{1-\rho_1^2\rho_2^2}{n(1-\rho_1^2)(1-\rho_2^2)}.$$

The ratio of the second of these quantities to the first (the asymptotic relative efficiency of the two tests) is

$$E(b_1, r \mid \rho=0) = \frac{1}{2}\frac{(1-\rho_1^2\rho_2^2)(1+\rho_1\rho_2)^2}{(1-\rho_1^4)(1-\rho_2^4)}.$$

which is evaluated for certain values of ρ_1 and ρ_2 in Table 3. For $E(b_1, r \mid \rho=0) \gg 1$ it may be inferred that, asymptotically at least, the test based on b_1 is superior to that based on r.

Table 3. $E(b_1, r \mid \rho=0)$

ρ_1 \ ρ_2	−0·8	−0·6	−0·4	−0·2	0	0·2	0·4	0·6	0·8
0	0·85	0·57	0·51	0·50	0·50	0·50	0·51	0·57	0·85
0·2	0·58	0·44	0·43	0·46	0·50	0·54	0·60	0·71	1·11
0·4	0·36	0·32	0·36	0·43	0·51	0·60	0·69	0·85	1·36
0·6	0·20	0·24	0·32	0·44	0·57	0·71	0·85	1·06	1·64
0·8	0·11	0·20	0·36	0·58	0·85	1·11	1·36	1·64	2·28

As in § 2 we have

$$\lim_{n\to\infty}\left\{\frac{\partial}{\partial\rho}\mathscr{E}(r(x_2y_2 \mid x_1))\right\}_{\rho=0} = \lim_{n\to\infty}\operatorname{cov}\{r(x_2y_2 \mid x_1), \mathbf{y}'\boldsymbol{\Gamma}_1^{-1}\boldsymbol{\Gamma}_3\boldsymbol{\Gamma}_2^{-1}\mathbf{x}\},$$

the differentiation having been carried out under the integral sign and $\mathbf{y}$ now being independent of $\mathbf{x}$. Here $\boldsymbol{\Gamma}_1$ is the covariance matrix of the y_t, $\boldsymbol{\Gamma}_2$ is the covariance matrix of the x_t and

$$\boldsymbol{\Gamma}_3 = \frac{1}{\rho}[\mathscr{E}\{(y_i-\mu_1)(x_j-\mu_2)\}].$$

The limit of the covariance can be evaluated in the same manner as in § 2 and gives

$$\lim_{n\to\infty}\left\{\frac{\partial}{\partial\rho}\mathscr{E}(r(x_2y_2 \mid x_1))\right\} = (1-\rho_1^2)^{\frac{1}{2}},$$

$$\lim_{n\to\infty}\left\{\frac{\partial}{\partial\rho}\mathscr{E}(r(x_2y_2 \mid y_1))\right\} = (1-\rho_2^2)^{\frac{1}{2}},$$

$$\lim_{n\to\infty}\left\{\frac{\partial}{\partial\rho}\mathscr{E}(r(x_2y_2 \mid x_1y_1))\right\} = 1.$$

Clearly once more $r(x_2y_2 \mid x_1y_1)$ provides the most efficient test statistic.

It can be shown that, for $\rho = 0$, the quantities $-\mathscr{E}\left(\frac{\partial^2 \log L}{\partial\rho\,\partial\theta_j}\right)_{\rho=0}$ are asymptotically zero. Here L is the likelihood function and the θ_j ($j = 1, 2, 3, 4$) are the remaining parameters in that function. At $\rho = 0$, $-\mathscr{E}\left(\frac{\partial^2 \log L}{\partial\rho^2}\right) = n$. It therefore appears (Whittle, 1953) that the asymptotic variance of the maximum-likelihood estimator of ρ is n^{-1} when ρ is zero. Since

this estimator will be consistent both r and b_1 are asymptotically inefficient as test statistics when ρ_1 and ρ_2 are not zero, while $r(x_2y_2 \mid x_1y_1)$ is asymptotically fully efficient.

When ρ_1 and ρ_2 are high and of the same sign b_1 will give an exact test of the null hypothesis which is more efficient than the test based on r. However, the test based on b_1 is always much less efficient than that based on $r(x_2y_2 \mid x_1y_1)$.

The computational burden of the exact test can of course be reduced by omitting one or both of x_{2l-1} and x_{2l+1} from the regression and the test will still be exact. It can be shown, however, that the test will be asymptotically less efficient. Another alternative is to use $(x_{2l-1}+x_{2l+1})$ in place of x_{2l-1} and x_{2l+1}. It can then be shown that the asymptotic efficiency of the test based on the coefficient of x_{2l} is unchanged. In finite samples, however, the power of the test may be adversely affected.

The following example illustrates the use of some of the tests mentioned in this section.

Example 2. In Moran (1952) the correlations between the records of the numbers of certain game birds shot over a period of years were considered. The series for caper and ptarmigan are considered in this example.

The application of the test given in Quenouille (1947*a*) shows that both series are well represented by a simple Markoff process.

The results of the various tests discussed above are shown in Table 4. Here r_m indicates a correlation coefficient with m degrees of freedom.

Table 4

Statistic	Sample value	Test statistic
r	0·417	$r_{34} = 0{\cdot}417^*$
b_1	0·444	$t_{31} = 2{\cdot}484^*$
$r(x_2y_2 \mid x_1y_1)$	0·387	$r_{69} = 0{\cdot}387^{***}$

* Significant at 5 % point. *** Significant at 0·1 % point.

The two observed serial correlations were 0·592 and 0·417 for caper and ptarmigan respectively, and these were adjusted up to 0·65 and 0·55 by Moran to allow for bias. Using these last values it appears from Table 3 that r and b_1 should give tests of roughly the same efficiency while $r(x_2y_2 \mid x_1y_1)$ should be approximately twice as efficient as either.

4. The case where the regressor process is a second-order autoregressive process

The exact tests for correlation between two series which have been presented in this paper remain exact provided that the residual from the regression of one variate on the other is a Markoff process. The efficiency of the tests has been examined in the following cases:

(1) The regressor process is a Markoff process wholly independent of the residual. In particular this covers the case where the regressand is also a Markoff process with the same parameter as the regressor and the residual.

(2) The regressand and regressor processes are both Markovian. When the two parameters are equal this case is equivalent to the particular case mentioned under (1) above.

If neither of the two processes being correlated can reasonably be said to be Markovian

then, strictly, the exact tests presented here are not applicable, for the residual can then be Markovian only if there is in truth a relation between the series. (One might apply the test as an approximation to an exact test of course.)

A case which may arise in practice is that in which one process appears Markovian while the other is an autoregressive process of higher order. The more interesting alternative hypothesis then appears to be one in which the correlation between the series arises from a correlation between the shocks of which they are linearly composed.

Let
$$(y_t-\mu_1) = \rho_1(y_{t-1}-\mu_1)+\epsilon_t$$
and
$$(x_t-\mu_2)+a(x_{t-1}-\mu_2)+b(x_{t-2}-\mu_2) = \eta_t$$
be two processes where,

(1) $|\rho_1| < 1$ and all roots of $z^2+az+b = 0$ are less than unity in absolute value,

(2) ϵ_t and η_t are jointly normally distributed with variances
$$\sigma_1^2(1-\rho_1^2) \quad \text{and} \quad \sigma_2^2\{(1-b)(1+b-a^2)(1+b)^{-1}\}$$
and correlation ρ.

Table 5. $E(b_1, r \mid \rho = 0)$

b \ ρ_1, ρ_2	$\rho_1=\rho_2=0{\cdot}4$	$\rho_1=\rho_2=0{\cdot}6$	$\rho_1=\rho_2=0{\cdot}8$	$\rho_1=0{\cdot}6$ $\rho_2=0{\cdot}8$	$\rho_1=0{\cdot}8$ $\rho_2=0{\cdot}6$
−0·6	0·82	1·42	3·45	2·73	1·82
−0·4	0·77	1·30	3·04	2·33	1·79
−0·2	0·73	1·18	2·65	1·97	1·73
0	0·69	1·06	2·28	1·64	1·64
0·2	0·65	0·95	1·94	1·36	1·53
0·4	0·62	0·86	1·64	1·12	1·42
0·6	0·59	0·79	1·40	0·95	1·33

Then by straightforward methods it can be shown that
$$E(b_1, r \mid \rho = 0) = \frac{(1+\rho_1\rho_2-b)^2\{1+\rho_1\rho_2-b(\rho_1^2+\rho_1\rho_2)\}\{1-\rho_1\rho_2-b(\rho_1\rho_2-\rho_1^2)\}}{2(1-\rho_1^4)(1-\rho_2^2)(1-b)\{1+\rho_2^2-b(1-\rho_2^2)\}}.$$
Here ρ_2 is the first serial correlation of the x_t process.

At $b = 0$ this is equal to the quantity tabulated above. The ratio decreases with b, for most values of ρ_1 and ρ_2, at least in the neighbourhood of $b = 0$. For sufficiently large values of ρ_1 and ρ_2 it will be greater than unity. The ratio is tabulated in Table 5 for certain values of ρ_1, ρ_2 and b.

It is clear that the statistic $r(x_3y_3 \mid x_1x_2y_2)$ will now give a test which is asymptotically fully efficient. The asymptotic efficiency of the statistic $r(x_2y_2 \mid x_1y_1)$ can be fairly easily evaluated and proves to be
$$E\{r(x_2y_2 \mid x_1y_1), r(x_3y_3 \mid x_1x_2y_2) \mid \rho = 0\} = (1-b^2).$$

The use of $r(x_2y_2 \mid x_1y_1)$ will then result in little loss of information unless b is fairly large.

Finally
$$E\{b_1, r(x_2y_2 \mid x_1y_1)\} = \frac{(1-b+\rho_1\rho_2)^2}{2(1+\rho_1^2)(1-b)\{1-b+\rho_2^2(1+b)\}}.$$

This is less than unity for most of the range of values of the parameters involved though it becomes infinite as b tends to unity.

Example 3. The data for grouse and blackgame from Moran (1952) may be used for illustration. The blackgame series appears to be Markovian but the first partial serial correlation for the grouse series is 0·34. The first serial correlation for grouse is 0·64 and for blackgame 0·43. Using $b = 0{\cdot}3$, $\rho_1 = 0{\cdot}7$ and $\rho_2 = 0{\cdot}5$ we obtain

$$E(b_1, r \mid \rho = 0) = 0{\cdot}02, \quad E\{b_1, r(x_2 y_2 \mid x_1 y_1 \mid \rho = 0\} = 0{\cdot}47.$$

The tests are shown in Table 6.

Table 6

Statistic	Sample value	Test statistic
b_1	0·014	$t_{31} = 2{\cdot}56$*
r	0·347	$r_{35} = 0{\cdot}347$*
$r(x_2 y_2 \mid x_1 y_1)$	0·327	$r_{69} = 0{\cdot}327$**

* Significant at 5% point. ** Significant at 1% point.

5. GENERALIZATION

The preceding exact tests can of course be generalized to multiple correlation and analogous results can be expected.

A further generalization of both the simple and multiple correlation tests can be obtained when one process is an autoregressive process of higher order than the first. For a process of order h however only $n(h+1)^{-1}$ observations will be used in the correlation (where n is the number of observations available), the remainder being used to reduce the series to independence in time. The serial correlations would have to be very high indeed before the test began to compare favourably with one based on the ordinary correlation (or multiple correlation) coefficient. The computational burden involved in the production of the partial correlation (or partial multiple correlation) coefficient will also be great.

6. SUMMARY

An exact test of the correlation between two series has been obtained which can be applied whenever one of the two series is Markovian. If the two series are x_t and y_t and y_t is Markovian the test statistic, b_1, is the partial correlation between x_{2t} and y_{2t} when the effects of $(y_{2t-1} + y_{2t+1})$, x_{2t-1} and x_{2t+1} have been removed. The asymptotic efficiency of this statistic is compared with that of the ordinary correlation coefficient, r, between the two series and the statistic $r(x_2 y_2 \mid x_1 y_1)$ suggested by Quenouille (1949), under the following conditions:

(*a*) The residual process from the regression of y_t on x_t is independent of the x_t process and comes from a Gaussian Markoff process.

(*b*) The two series x_t and y_t are Markovian and are correlated through a correlation between the (Gaussian) errors in the two processes.

(*c*) As in (*b*) but with x_t generated by a second order autoregressive process.

While in all three cases the statistic b_1 leads to an asymptotically more efficient test than r for high serial correlation of the residual process in the regression of y_t on x_t, the statistic

$r(x_2y_2 \mid x_1y_1)$ is always asymptotically more efficient than b_1 with the exception of some cases, under (*c*), where the first partial correlation of the x_t process is high and positive.

I should like to thank a referee of this paper for pointing out the importance of Quenouille's statistic.

REFERENCES

BARTLETT, M. S. (1935). *J. R. Statist. Soc.* **98**, 536.
COCHRANE, D. & ORCUTT, G. H. (1949). *J. Amer. Statist. Ass.* **44**, 32.
CRAMER, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.
DOOB, J. L. (1953). *Stochastic Processes*. New York: John Wiley and Sons.
HANNAN, E. J. (1955). *Biometrika*, **42**, 133.
KENDALL, M. G. (1949). *Biometrika*, **36**, 267.
MOOD, A. M. (1954). *Ann. Math. Statist.* **25**, 514.
MORAN, P. A. P. (1952). *J. Animal Ecol.* **21**, 154.
OGAWARA, M. (1951). *Ann. Math. Statist.* **22**, 115.
PITMAN, E. J. G. (1948). Lecture notes on non-parametric inference. University of North Carolina.
QUENOUILLE, M. H. (1947*a*). *J. R. Statist. Soc.* **60**, 123.
QUENOUILLE, M. H. (1947*b*). *Biometrika*, **34**, 365.
QUENOUILLE, M. H. (1949). *J. R. Statist. Soc.* B, **11**, 68.
RAO, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. New York: John Wiley and Sons.
STUART, A. (1954). *J. Amer. Statist. Ass.* **49**, 147.
WHITTLE, P. (1953). *J. R. Statist. Soc.* B, **15**, 125.
WOLD, H. (1953). *Demand Analysis*. New York: John Wiley and Sons.

SERIAL CORRELATION IN REGRESSION ANALYSIS. I†

BY G. S. WATSON

Australian National University, Canberra, A.C.T.

1. INTRODUCTION

The procedure for linear regression analysis when the errors have a multivariate normal joint distribution with covariance matrix $\sigma^2\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is a *given* matrix are well known (see, for example, Aitken, 1934–5). However, in many practical applications, the matrix $\boldsymbol{\alpha}$ is not completely known. In the analysis of ordered observations, the errors are often serially correlated. Even if the errors are assumed to be generated by a stationary stochastic process of the moving-average or autoregressive type of lower order, the actual selection of the process and its parameters is a difficult matter, requiring samples of a size not often available. It is therefore certain that regression analyses will often be made using the wrong matrix $\boldsymbol{\alpha}$.

Recently Hannan (1955) has given a method which, while not fully efficient, leads to exact tests when the error process is Markov with unknown ρ. This is an advance on previous writers (e.g. Wold, 1949; Cochrane & Orcutt, 1949), who aimed at the estimation of the serial correlations of the errors and hence the determination of $\boldsymbol{\alpha}$. But the basic difficulty still remains—the error process may not be Markov. Furthermore, while the difficulties are well known, the orders of magnitudes of the effects of using the wrong matrix $\boldsymbol{\alpha}$ have received little precise study except for the case of unequal variances in the analysis of variance (see, for example, Box, 1954).

For these reasons in Part I of this paper an examination will be made of the performance of a regression analysis based on the assumption that the error covariance matrix is $\sigma^2\boldsymbol{\gamma}$ when it is, in fact, $\sigma^2\boldsymbol{\alpha}$. The methods used have their origin in the papers of Durbin & Watson (1950, 1951). When $\boldsymbol{\gamma} \neq \boldsymbol{\alpha}$ the regression vectors play an important role in determining the efficiency of the regression coefficient estimates and the true significance levels of significance tests. The procedure adopted is to find, for fixed $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, the regression vectors which would make the analysis as 'bad' as possible; this leads to inequalities on the bias in the estimates of variance of the regression coefficients, on the efficiency of the estimates of the regression coefficients and on the significance points of the various t- and F-tests.

In Part II, these inequalities will be applied to cases where $\sigma^2\boldsymbol{\alpha}$ and $\sigma^2\boldsymbol{\gamma}$ are the covariance matrices of first- and second-order autoregressive and moving-average processes. By using approximations to these matrices, it is possible to get simple algebraic and numerical results.

We consider the linear regression of a dependent variable y on k (linearly independent) regression variables $x_1, x_2, \ldots, x_k$. The model for a sample of N observations is, in matrix form,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{k1} \\ \vdots & & \vdots \\ x_{1N} & \ldots & x_{kN} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix}, \tag{1·1}$$

† This work was carried out while the writer was a Research Officer in the Department of Applied Economics, University of Cambridge. The results of this paper and its sequel were presented at the Conference of the Econometric Society, Louvain, Belgium in September 1951. It formed part of a thesis which was issued by the Institute of Statistics, University of North Carolina, Mimeograph Series, no. 49.

or $$\mathbf{y} = [\mathbf{x}_1, \ldots, \mathbf{x}_k]\boldsymbol{\beta} + \mathbf{u},$$

or $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where the error vector $\mathbf{u}$ has zero expectation and

$$E(\mathbf{u}\mathbf{u}') = \sigma^2\boldsymbol{\alpha} \quad (\boldsymbol{\alpha} \text{ non-singular}). \tag{1·2}$$

σ^2, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are supposed unknown. If the statistician, analysing these data, decides *a priori* that the error covariance matrix is $\sigma^2\boldsymbol{\gamma}$, he will essentially proceed as follows. Since $\boldsymbol{\gamma}$ is known it is possible to find a non-singular $N \times N$ matrix $\mathbf{H}$ such that

$$\mathbf{H}\boldsymbol{\gamma}\mathbf{H}' = \mathbf{I}. \tag{1·3}$$

Writing $$\mathbf{H}\mathbf{y} = \mathbf{y}^*, \quad \mathbf{H}\mathbf{x}_i = \mathbf{x}_i^*, \quad \mathbf{H}\mathbf{u} = \mathbf{u}^*, \quad \mathbf{H}\mathbf{X} = \mathbf{X}^*,$$

the result of applying $\mathbf{H}$ to equation (1·1) is the transformed regression equation

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^*, \tag{1·4}$$

where $\mathbf{u}^*$ has zero expectation and

$$E(\mathbf{u}^*\mathbf{u}'^*) = \sigma^2\mathbf{H}\boldsymbol{\alpha}\mathbf{H}'. \tag{1·5}$$

The regression equation (1·4) will now be treated by the least-squares procedures.† Thus in our analysis of the effects of 'guessing' the error covariance matrix, we may without loss of generality proceed on the assumption that the least-squares procedure has been used on the model represented by equations (1·1) and (1·2) and merely replace $\boldsymbol{\alpha}$ by $\mathbf{H}\boldsymbol{\alpha}\mathbf{H}'$ in the final results.

On these assumptions, the estimate of $\boldsymbol{\beta}$ used would be

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{1·6}$$

The covariance matrix of the components of $\mathbf{b}$ is actually

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\alpha}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \tag{1·7}$$

but it would be taken as

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \tag{1·8}$$

where σ^2 would be estimated by

$$s^2 = \frac{(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})}{N-k}. \tag{1·9}$$

The true expected value of s^2 is given by

$$\begin{aligned} E(s^2) &= \frac{1}{N-k}E((\mathbf{u}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})'(\mathbf{u}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})) \\ &= \frac{1}{N-k}E(\mathbf{u}'\mathbf{u}-\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}) \\ &= \frac{\sigma^2}{N-k}(\operatorname{tr}\boldsymbol{\alpha}-\operatorname{tr}\mathbf{X}'\boldsymbol{\alpha}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}). \end{aligned} \tag{1·10}$$

For arbitrary $\mathbf{X}$ this reduces to σ^2 only when $\boldsymbol{\alpha} = \mathbf{I}$.

† By the 'least-squares procedures' is meant not only the estimation of the regression coefficients by least squares but also the use of the variance estimates and the t- and F-tests which are appropriate to the case where the error covariance matrix really is $\mathbf{I}\sigma^2$. We might have introduced the name 'I-procedure' to cover this. Thus when the statistician decides the error covariance matrix is $\sigma^2\boldsymbol{\gamma}$ he will use the '$\boldsymbol{\gamma}$-procedure'.

2. BOUNDS ON THE BIAS OF THE VARIANCE ESTIMATES

The bias in the estimate of the covariance matrix of the regression coefficient estimates may be found from (1·7), (1·8) and (1·10) and is given by

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\left(\frac{\operatorname{tr}\boldsymbol{\alpha}-\operatorname{tr}\mathbf{X}'\boldsymbol{\alpha}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}}{N-k}\mathbf{I}-\mathbf{X}'\boldsymbol{\alpha}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right). \tag{2·1}$$

In order to examine the magnitude of the effect on the variance of a single regression coefficient we now assume that the regression vectors form an orthonormal set, i.e.

$$\mathbf{X}'\mathbf{X}=\mathbf{I}. \tag{2·2}$$

The condition (2·2) invoked to simplify (2·1) leads to a loss of generality; but, without it, the extreme value problem is very difficult. Fortunately we will not be so restricted when we come to discuss significance tests. On this assumption the estimate of β_i will be

$$b_i=\mathbf{x}_i'\mathbf{y}, \tag{2·3}$$

and the bias in the estimate of its variance will be

$$\sigma^2\left(\frac{\operatorname{tr}\boldsymbol{\alpha}-\operatorname{tr}\mathbf{X}'\boldsymbol{\alpha}\mathbf{X}}{N-k}-\mathbf{x}_i'\boldsymbol{\alpha}\mathbf{x}_i\right),$$

that is,

$$\sigma^2\left(\frac{\operatorname{tr}\boldsymbol{\alpha}-\sum_{j=1}^{k}\mathbf{x}_j'\boldsymbol{\alpha}\mathbf{x}_j}{N-k}-\mathbf{x}_i'\boldsymbol{\alpha}\mathbf{x}_i\right). \tag{2·4}$$

We now require the extremes of the expression (2·4). These will be seen to follow easily from the following algebraic result: the extremes of the quadratic form $\mathbf{Z}'\boldsymbol{\alpha}\mathbf{Z}$, where $\mathbf{Z}$ is a unit vector in a subspace spanned by a subset of the latent vectors of $\boldsymbol{\alpha}$, are the least and greatest latent roots of $\boldsymbol{\alpha}$ associated with the subspace. Thus the lower bound of (2·4) is obtained by choosing one of the x-vectors, say $\mathbf{x}_i$, to be the latent vector of $\boldsymbol{\alpha}$ corresponding to its largest latent root and the others to be the latent vectors corresponding to the next $k-1$ largest roots. The upper bound is found by a similar choice using the smallest roots of $\boldsymbol{\alpha}$. More generally if $h\ (<k)$ of the regression vectors are latent vectors of $\boldsymbol{\alpha}$, the remaining regression vectors lie in an $N-h$ dimensional subspace. If $\mathbf{x}_i$ lies in this subspace, the above results hold good if the choices are made with respect to the set of the $N-h$ latent roots associated with this subspace. If $\mathbf{x}_i$ is a latent vector, a trivial modification is necessary.

Thus for $\mathbf{X}$ subject only to (2·2)

$$\left.\begin{aligned}&(a)\ \text{Maximum bias} = \sigma^2((\text{mean of } N-k \text{ greatest roots of } \boldsymbol{\alpha})-(\text{least root of } \boldsymbol{\alpha})),\\ &(b)\ \text{Minimum bias} = \sigma^2((\text{mean of } N-k \text{ least roots of } \boldsymbol{\alpha})-(\text{greatest root of } \boldsymbol{\alpha})).\end{aligned}\right\} \tag{2·5}$$

If the multiple σ^2 is omitted the results are in the nature of bounds on the fractional bias. It will be noticed that (2·5 (*a*)) is positive and (2·5 (*b*)) negative *whatever* $\boldsymbol{\alpha}$ may be. It is commonly believed that presence of serial correlation makes the variance estimates deceptively small. The numerical results given in Part II show this to be the stronger tendency, but (2·5) shows that this need not always be the case. The deciding factor is seen to be the relationship of the regression vectors to the latent vectors of $\boldsymbol{\alpha}$. Finally, (2·5) shows how the bias must tend to zero as $\boldsymbol{\alpha}\to\mathbf{I}$.

3. Lower bound to the efficiency of the estimates

We turn now to a consideration of the efficiency of the estimates (1·6) of the regression coefficients. It is necessary to determine how this efficiency falls off as $\boldsymbol{\alpha}$ differs more and more from $\mathbf{I}$. To consider the efficiency of the estimation procedure as a whole it is convenient to introduce the generalized variance of the estimates. The generalized variance of the estimates b_i of (1·6) is defined to be the determinant of their covariance matrix (1·7). Aitken (1948) has shown that the linear estimates of β_i with the least generalized variance are those given by $(\mathbf{X}'\alpha^{-1}\mathbf{X})^{-1}\mathbf{X}'\alpha^{-1}\mathbf{y}$; they have the covariance matrix $\sigma^2(\mathbf{X}'\alpha^{-1}\mathbf{X})^{-1}$. The efficiency of the estimates (1·6) is then naturally defined as the ratio of the determinants of the covariance matrices $\sigma^2(\mathbf{X}'\boldsymbol{\alpha}^{-1}\mathbf{X})^{-1}$ and (1·7). Thus

$$\text{Eff.}(\mathbf{b}) = \frac{|\mathbf{X}'\mathbf{X}|^2}{|\mathbf{X}'\boldsymbol{\alpha}\mathbf{X}|\,|\mathbf{X}'\boldsymbol{\alpha}^{-1}\mathbf{X}|}. \tag{3·1}$$

It should be noted that Eff. $(\mathbf{b})$ is invariant when $\mathbf{X}$ is transformed to $\mathbf{P}$ by $\mathbf{P} = \mathbf{XG}$. It is always possible to find $\mathbf{G}$ so that $\mathbf{P}'\mathbf{P} = \mathbf{I}$. Thus it may be assumed without loss of generality that $\mathbf{X}$ satisfies (2·2).

If all the regression vectors are latent vectors† of $\boldsymbol{\alpha}$, Eff. $(\mathbf{b}) = 1$. This may be verified immediately since only the diagonal terms in $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\boldsymbol{\alpha}\mathbf{X}$ and $\mathbf{X}'\boldsymbol{\alpha}^{-1}\mathbf{X}$ are then non-zero. The result finds an application in the work of R. L. & T. W. Anderson (1950) on testing for circular serial correlation when a short Fourier series has been fitted by least squares. In this problem the regression vectors are the latent vectors of the quadratic form occurring in the joint density function of successive observations from a circular autoregressive process so that the estimates of the regression coefficients are optimal whether serial correlation, of the type envisaged, is present or not.‡

The expression (3·1) tends of course to unity as $\boldsymbol{\alpha} \to \mathbf{I}$. The important question here is how small Eff. $(\mathbf{b})$ may become, where $\boldsymbol{\alpha}$ is fixed and not equal to $\mathbf{I}$ but when $\mathbf{X}$ is unrestricted. This lower bound to the efficiency may be obtained by generalizing an inequality due§ to J. W. S. Cassels. Cassels's inequality states that for fixed $a_j > 0$, $b_j > 0$ and all $w_j \geqslant 0$ $(j = 1, \ldots, N)$

$$\frac{4rR}{(r+R)^2} \leqslant \frac{\left(\sum_1^N a_j b_j w_j\right)^2}{\left(\sum_1^N a_j^2 w_j\right)\left(\sum_1^N b_j^2 w_j\right)} \leqslant 1, \tag{3·2}$$

where $r = \min_j a_j/b_j$, $R = \max_j a_j/b_j$. If the numbering of the a_j and b_j is so arranged that $r = a_1/b_1$ and $R = a_N/b_N$, then the lower bound is attained for $w_1 = 1/a_1 b_1$ and $w_N = 1/a_N b_N$, $w_j = 0$ $(j \neq 0, N)$. The lower bound in (3·2) is the square of the ratio of the geometric to the arithmetic mean of r and R.

To obtain the minimum of Eff. $(\mathbf{b})$ by varying $\mathbf{X}$ subject to the restriction that the regression vectors are orthogonal, we first eliminate from the problem any regression

† In contrast to this, the bias matrix (2·1) is not zero when the regression vectors are latent vectors of $\boldsymbol{\alpha}$.

‡ This result is, however, little more than a mathematical curiosity. For apart from the fact that the Andersons' test has power against other alternatives, such estimates while optimal cannot be validly tested in the usual way.

§ A proof of this inequality is given in the Appendix. The inequality and the proof were communicated privately by Dr Cassels to the author, who wishes to record his gratitude.

vectors which are latent vectors of $\boldsymbol{\alpha}$. Suppose that $\mathbf{x}_{k-h+1}, \ldots, \mathbf{x}_k$ are latent vectors of $\boldsymbol{\alpha}$ associated with the latent roots $\alpha_{N-h+1}, \ldots, \alpha_N$. If the remaining $\mathbf{x}_i$ have co-ordinates Z_{it} $(t = 1, \ldots, N)$ when referred to the latent vectors of $\boldsymbol{\alpha}$ as axes

$$\text{Eff. } (\mathbf{b}) = \frac{\prod_{i=1}^{k-h} \left(\sum_{t=1}^{N-h} Z_{it}^2 \right)^2}{\left| \sum_{t=1}^{N-h} \alpha_t Z_{it} Z_{jt} \right| \left| \sum_{t=1}^{N-h} \frac{1}{\alpha_t} Z_{it} Z_{jt} \right|}. \tag{3·3}$$

The matrices $\mathbf{X}'\boldsymbol{\alpha}\mathbf{X}$ and $\mathbf{X}'\boldsymbol{\alpha}^{-1}\mathbf{X}$ of (3·1), and therefore the matrices in the denominator of (3·3), are positive-definite by their form. By a theorem of Hadamard (see Hardy, Littlewood & Polya, 1934, p. 34) if c_{ij} are the elements of positive-definite matrix

$$|\mathbf{c}_{ij}| \leqslant c_{11} c_{22} \ldots c_{NN}. \tag{3·4}$$

Thus

$$\text{Eff. } (\mathbf{b}) \geqslant \prod_{i=1}^{k-h} \frac{\left(\sum_{t=1}^{N-h} Z_{it}^2 \right)^2}{\left(\sum_{t=1}^{N-h} \alpha_t Z_{it}^2 \right) \left(\sum_{t=1}^{N-h} \frac{1}{\alpha_t} Z_{it}^2 \right)}.$$

Without loss of generality we may assume that $\alpha_1 \leqslant \alpha_2 \leqslant \ldots \leqslant \alpha_{N-h}$. We now choose the $\mathbf{Z}$ vectors to minimize the right-hand side of (3·4) applying Cassels's inequality successively. When this is done, the matrix $\mathbf{X}$ takes the form

$$\left[\frac{\mathbf{a}_1 + \mathbf{a}_{N-h}}{\sqrt{2}}, \frac{\mathbf{a}_2 + \mathbf{a}_{N-h-1}}{\sqrt{2}}, \ldots, \frac{\mathbf{a}_{k-h} + \mathbf{a}_{N-k+1}}{\sqrt{2}}, \mathbf{a}_{N-h+1}, \ldots, \mathbf{a}_N \right].$$

Using (3·2) with $a_t = \sqrt{\alpha_t}, a_t b_t = 1$ and $w_{it} = Z_{it}^2$, we have the result†

$$\frac{4\alpha_1 \alpha_{N-h}}{(\alpha_1 + \alpha_{N-h})^2} \frac{4\alpha_2 \alpha_{N-h-1}}{(\alpha_2 + \alpha_{N-h-1})^2} \ldots \frac{4\alpha_{k-h} \alpha_{N-k+1}}{(\alpha_{k-h} + \alpha_{N-k+1})^2} \leqslant \text{Eff. } (\mathbf{b}) \leqslant 1. \tag{3·5}$$

Thus in spite of the use of Hadamard's inequality, we have arrived at a lower bound for Eff. $(\mathbf{b})$ which is attainable. This is so because we were led to a choice of $\mathbf{X}$ for which $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\boldsymbol{\alpha}\mathbf{X}$ and $\mathbf{X}'\boldsymbol{\alpha}^{-1}\mathbf{X}$ are diagonal. The general rule for the formation of the lower bound to the efficiency is evident. From the $N-h$ roots of $\boldsymbol{\alpha}$ not associated with any regression vectors, we choose successively the $(k-h)$ most extreme pairs. The lower bound to the efficiency thus depends on the ratios of the extreme roots of $\boldsymbol{\alpha}$. When the roots of $\boldsymbol{\alpha}$ are all approximately equal (approximately equal to unity if the error process is stationary) the efficiency must be high. It will be shown in Part II that high serial correlation of the autoregressive and moving average types leads to $\boldsymbol{\alpha}$ having roots of very different magnitudes and therefore to the possibility of low efficiencies. Finally, it will be noted that the maximum and minimum efficiencies are taken when the arbitrary regression vectors are respectively latent vectors of $\boldsymbol{\alpha}$ and sums of certain pairs of latent vectors of $\boldsymbol{\alpha}$; using the invariance properties of (3·1) these statements serve to define the maximizing and minimizing regression spaces.

4. Bounds on the significance points of the t- and F-tests

The significance tests usually made in regression analysis relate to linear hypotheses about the regression coefficients. Thus in the model (1·1), these hypotheses will be special cases of a null hypothesis specifying the value of one or more linear functions of $\beta_1, \ldots, \beta_k$. When the

† This form of the lower bound was conjectured by J. Durbin.

errors are independently distributed with the same normal distribution, it is well known that the optimal test statistics for such hypotheses have the t- or F-distributions. In this section bounds will be obtained for the distributions of these statistics when the covariance matrix of the errors is $\sigma^2\boldsymbol{\alpha}$; these bounding distributions converge to the tabulated distributions as $\boldsymbol{\alpha} \to \mathbf{I}$. We begin by showing that there is no loss of generality in considering only hypotheses specifying the value of several regression coefficients and in assuming that the regression vectors form an orthonormal set. The bounding distributions for this reduced problem will then be derived. The section concludes with a short discussion of the numerical determination of the bounding significance points.

We consider again the model (1·1), $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where the errors $\mathbf{u}$ have a multivariate-normal distribution with zero mean vector and covariance matrix $\sigma^2\boldsymbol{\alpha}$ and where the regression vectors are linearly independent. We suppose that the null hypothesis to be tested is

$$\mathbf{f}_i'\boldsymbol{\beta} = \phi_i \quad (i = 1, \ldots, h \leqslant k), \tag{4·1}$$

where $\mathbf{f}_i, \phi_i\,(i = 1, \ldots, h)$ are given and $\mathbf{f}_i\,(i = 1, \ldots, h)$ are linearly independent. The optimal test of the hypothesis (4·1) when $\boldsymbol{\alpha} = \mathbf{I}$ is obtained by building up a quadratic form in the variables $\mathbf{f}_i'\mathbf{b} - \phi_i$ ($\mathbf{b}$ is the least-squares estimate of $\boldsymbol{\beta}$ defined in (1·6)) which is distributed as a multiple of χ^2 with h degrees of freedom. The form must therefore be based on the inverse of the covariance matrix of the variables $\mathbf{f}_i'\mathbf{b} - \phi_i$. Because σ^2 is unknown it is necessary to divide this form by the estimate s^2 of σ^2 given in (1·9) in order to obtain a test statistic. The ratio so obtained has the F-distribution with h and $N-k$ degrees of freedom provided the null hypothesis is true. This is the statistic we must examine when $\boldsymbol{\alpha}$ is not necessarily equal to $\mathbf{I}$.

If $\mathbf{X}$ is transformed to $\mathbf{Z}$ by the transformation $\mathbf{P}$

$$\mathbf{Z} = \mathbf{X}\mathbf{P}, \tag{4·2}$$

it is always possible to find a non-singular $\mathbf{P}$ so that $\mathbf{Z}'\mathbf{Z} = \mathbf{I}$. For $\mathbf{P}$ has only to satisfy

$$\mathbf{P}'\mathbf{X}'\mathbf{X}\mathbf{P} = \mathbf{I}. \tag{4·3}$$

Defining $\boldsymbol{\delta} = \mathbf{P}^{-1}\boldsymbol{\beta}$, where $\mathbf{P}$ satisfies (4·3), (1·1) becomes

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\delta} + \mathbf{u}. \tag{4·4}$$

The least-squares estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, $\mathbf{b}$ and $\mathbf{d}$ say, are given by $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, $\mathbf{d} = \mathbf{Z}'\mathbf{y}$. It is easily seen that

$$\mathbf{d} = \mathbf{P}^{-1}\mathbf{b}. \tag{4·5}$$

The denominator of the test statistic is s^2, and it has been seen in the derivation of (1·10) that s^2 is proportional to

$$\mathbf{u}'\mathbf{u} - \mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{u}, \tag{4·6}$$

which becomes, under the transformation (4·2) and (4·3),

$$\mathbf{u}'\mathbf{u} - \mathbf{u}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u} = \mathbf{u}'\mathbf{u} - \mathbf{u}'\mathbf{Z}\mathbf{Z}'\mathbf{u}. \tag{4·7}$$

Thus (4·6) is invariant under transformations (4·2) and (4·7) is invariant under transformations (4·2) with $\mathbf{P}\mathbf{P}' = \mathbf{I}$. On the null hypothesis, the variables in the numerator are given by

$$\mathbf{f}_i'\mathbf{b} - \phi_i = \mathbf{f}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \quad (i = 1, \ldots, h). \tag{4·8}$$

Thus for $i = 1, \ldots, h$, $\mathbf{f}_i'\mathbf{b} - \phi_i$ are distributed as $\mathbf{f}_i'\mathbf{b}$ when $\boldsymbol{\beta} = 0$ or as $\mathbf{f}_i'\mathbf{P}\mathbf{d}$ when $\boldsymbol{\delta} = 0$. We may express the numerator quadratic form in terms of $\mathbf{d}$ without first stating it in terms of $\mathbf{b}$. Writing $\mathbf{f}_i'\mathbf{P} = \mathbf{g}_i'$, it must on the null hypothesis have the form

$$[\mathbf{g}_1'\mathbf{d}, \ldots, \mathbf{g}_h'\mathbf{d}] \begin{bmatrix} \mathbf{g}_1'\mathbf{g}_1 & \ldots & \mathbf{g}_1'\mathbf{g}_h \\ \vdots & & \vdots \\ \mathbf{g}_h'\mathbf{g}_1 & \ldots & \mathbf{g}_h'\mathbf{g}_h \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{g}_1'\mathbf{d} \\ \vdots \\ \mathbf{g}_h'\mathbf{d} \end{bmatrix}, \tag{4·9}$$

because this is seen to be correct when $\boldsymbol{\alpha} = \mathbf{I}$. (4·9) may be written as

$$\mathbf{d}'[\mathbf{g}_1, \ldots, \mathbf{g}_h] \begin{bmatrix} \mathbf{g}_1'\mathbf{g}_1 & \ldots & \mathbf{g}_1'\mathbf{d}_h \\ \vdots & & \vdots \\ \mathbf{g}_h'\mathbf{g}_1 & \ldots & \mathbf{g}_h'\mathbf{g}_h \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{g}_1' \\ \vdots \\ \mathbf{g}_h' \end{bmatrix} \mathbf{d}. \tag{4·10}$$

The matrix of this quadratic form in $\mathbf{d}$ is symmetric and has h unit roots and $k-h$ zero roots. Thus an orthogonal transformation $\mathbf{H}$ may be found so that if $\mathbf{d} = \mathbf{H}\mathbf{c}$, (4·10) reduces to $\mathbf{c}'\mathbf{J}\mathbf{c}$, where

$$\mathbf{J} = \begin{bmatrix} \mathbf{I}_h & \mathbf{0}_1 \\ \mathbf{0}_2 & \mathbf{0}_3 \end{bmatrix},$$

a $k \times k$ matrix with units in the first h places on the leading diagonal and zeros elsewhere. Since $\mathbf{H}$ is orthogonal,

$$\mathbf{c} = \mathbf{H}^{-1}\mathbf{d} = \mathbf{H}'\mathbf{Z}'\mathbf{u}. \tag{4·11}$$

If finally we put

$$\mathbf{W} = \mathbf{Z}\mathbf{H}, \tag{4·12}$$

then

$$\mathbf{c} = \mathbf{W}'\mathbf{u},$$

and the test statistic is proportional to

$$\frac{\mathbf{u}'\mathbf{W}\mathbf{J}\mathbf{W}'\mathbf{u}}{\mathbf{u}'\mathbf{u} - \mathbf{u}'\mathbf{W}\mathbf{W}'\mathbf{u}} = \frac{(\mathbf{u}'\mathbf{w}_1)^2 + \ldots + (\mathbf{u}'\mathbf{w}_h)^2}{\mathbf{u}'\mathbf{u} - (\mathbf{u}'\mathbf{w}_1)^2 - \ldots - (\mathbf{u}'\mathbf{w}_h)^2}, \tag{4·13}$$

where $\mathbf{w}_1, \ldots, \mathbf{w}_k$ form an orthonormal set of vectors. This is the desired canonical form for the test statistic because it is the form which arises in testing the hypothesis

$$\psi_i = \psi_i^0 \quad (i = 1, \ldots, h)$$

in the model

$$\mathbf{y} = \mathbf{W}\boldsymbol{\psi} + \mathbf{u},$$

where $\mathbf{W}'\mathbf{W} = \mathbf{I}$ and $\boldsymbol{\psi} = \{\psi_1, \ldots, \psi_k\}$. $\mathbf{c}$ defined above is the least-squares estimate of $\boldsymbol{\psi}$.

Since the proof above for the general linear hypothesis presents some difficulty, it seems worth while to establish the result for a simple particular case. Suppose that the null hypothesis is

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0. \tag{4·14}$$

To obtain the form of the test statistic, we refer to (1·8) which defines the covariance matrix of $\mathbf{b} - \boldsymbol{\beta}_0$ and (1·9) which gives an expression for s^2. Thus the statistic is proportional to

$$\frac{(\mathbf{b} - \boldsymbol{\beta}_0)'(\mathbf{X}'\mathbf{X})(\mathbf{b} - \boldsymbol{\beta}_0)}{(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})}. \tag{4·15}$$

Since $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, the statistic (4·15) on the null hypothesis may be put in the form

$$\frac{\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}}{\mathbf{u}'\mathbf{u} - \mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}}. \tag{4·16}$$

If now $\mathbf{X}$ is transformed to $\mathbf{Z}$, so that $\mathbf{Z}'\mathbf{Z} = \mathbf{I}$, the discussion of (4·2) ... (4·7) shows that (4·16) is reduced to

$$\frac{\mathbf{u}'\mathbf{Z}\mathbf{Z}'\mathbf{u}}{\mathbf{u}'\mathbf{u} - \mathbf{u}'\mathbf{Z}\mathbf{Z}'\mathbf{u}}. \tag{4·17}$$

This is the required canonical form for the hypothesis (4·14).

We consider the model of (1·1) and (1·2) under the assumptions of orthonormality (2·2). The errors $\mathbf{u}$ will now be assumed to have a multivariate normal distribution with zero mean vector and covariance matrix $\sigma^2\boldsymbol{\alpha}$. We suppose initially that the first $k-1$ of the regression vectors $\mathbf{x}$ are latent vectors of $\boldsymbol{\alpha}$, the remaining vector being arbitrary. Denote $\mathbf{x}_i = \mathbf{a}_i$ where $\boldsymbol{\alpha}\mathbf{a}_i = \alpha_i \mathbf{a}_i$ $(i = 1, \ldots, k-1)$. To test the hypothesis

$$\beta_k = \beta_k^0 \tag{4·18}$$

the appropriate statistic is

$$t^2 = \frac{(N-k)(b_k - \beta_k^0)^2}{\sum_{l=1}^{N} \left(y_l - \sum_{1}^{k-1} b_i a_{il} - b_k x_{kl}\right)^2}. \tag{4·19}$$

On the null hypothesis $t^2/(N-k)$ has the matrix form

$$\frac{t^2}{N-k} = \frac{\mathbf{u}'\mathbf{x}_k\mathbf{x}_k'\mathbf{u}}{\mathbf{u}'(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1' - \ldots - \mathbf{a}_{k-1}\mathbf{a}_{k-1}' - \mathbf{x}_k\mathbf{x}_k')\mathbf{u}} = T, \quad \text{say.} \tag{4·20}$$

Bounds on the significance points of T will now be found which enclose the true significance point of T, no matter what the direction of the vector $\mathbf{x}_k$.

By a device, often used in the treatment of ratios, we have

$$P(T < \tau) = P(\mathbf{u}'(\mathbf{x}_k\mathbf{x}_k' - \tau(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1' - \ldots - \mathbf{a}_{k-1}\mathbf{a}_{k-1}' - \mathbf{x}_k\mathbf{x}_k'))\mathbf{u} < 0). \tag{4·21}$$

But the quadratic form on the right-hand side of (4·21) is distributed like $\sum_{i=1}^{N} \nu_i v_i^2$, where ν_i are the roots of the determinantal equation

$$|\mathbf{x}_k\mathbf{x}_k' - \tau(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1' - \ldots - \mathbf{a}_{k-1}\mathbf{a}_{k-1}' - \mathbf{x}_k\mathbf{x}_k') - \nu\boldsymbol{\alpha}^{-1}| = 0, \tag{4·22}$$

and the v_i are N.I.D. $(0, 1)$. For fixed τ, let $\mathbf{x}_k$ vary. Then the ν_i and $P(\Sigma\nu_i v_i^2 < 0)$ vary. We will show that there exists a direction of $\mathbf{x}_k$ such that *all* the ν_i take maximum values. For this $\mathbf{x}_k$, $P(\Sigma\nu_i v_i^2 < 0)$ is clearly a minimum. Since, for fixed $\mathbf{x}_k$, $P(T < \tau)$ is a monotonically increasing function of τ, it follows that the significance point of T, τ_U say, for this special maximizing direction is greater than the significance point at the same level for any other vector $\mathbf{x}_k$. In the same way a lower bound τ_L will be established.

For this purpose apply to (4·22) the orthogonal transformation $\mathbf{K}$ which diagonalizes $\boldsymbol{\alpha}$. If

$$\left.\begin{aligned} \mathbf{K}'\mathbf{x}_k &= \mathbf{m}, \\ \mathbf{K}'\boldsymbol{\alpha}^{-1}\mathbf{K} &= \begin{bmatrix} 1/\alpha_1 & & & \\ & 1/\alpha_2 & & \\ & & \ddots & \\ & & & 1/\alpha_N \end{bmatrix}, \end{aligned}\right\} \tag{4·23}$$

then the first $k-1$ co-ordinates of $\mathbf{m}$ are zero because $\mathbf{x}_k$ is required to be orthogonal to the $k-1$ latent vectors corresponding to $\alpha_1, \ldots, \alpha_{k-1}$, and

$$\mathbf{K}'(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1' - \ldots - \mathbf{a}_{k-1}\mathbf{a}_{k-1}')\mathbf{K} = \begin{bmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & & 1 & \\ & & & & \ddots \\ & & & & & 1 \end{bmatrix}, \tag{4·24}$$

where the matrix on the right-hand side of (4·24) is an $N \times N$ identity matrix with the first $k-1$ diagonal elements replaced by zeros. Thus equation (4·22) has $k-1$ zero roots and $N-k+1$ which satisfy the reduced determinantal equation

$$\begin{vmatrix} (1+\tau)m_k^2-\tau-\dfrac{\nu}{\alpha_k}, & (1+\tau)m_k m_{k+1}, & \dots, & (1+\tau)m_k m_N \\ (1+\tau)m_{k+1}m_k, & (1+\tau)m_{k+1}^2-\tau-\dfrac{\nu}{\alpha_{k+1}}, & \dots, & (1+\tau)m_{k+1}m_N \\ \vdots & \vdots & & \vdots \\ (1+\tau)m_N m_k, & (1+\tau)m_N m_{k+1}, & \dots, & (1+\tau)m_N^2-\tau\dfrac{\nu}{\alpha_N} \end{vmatrix} = 0. \quad (4\cdot25)$$

Subtracting m_{k+1}/m_k times the first from the second row, etc., and expanding the result as a bordered determinant we find the equation in ν,

$$\prod_{p=k}^{N}\left(\tau+\frac{\nu}{\alpha_p}\right)-(1+\tau)\sum_{p=k}^{N}m_p^2\prod_{j\neq p}\left(\tau+\frac{\nu}{\alpha_j}\right)=0, \quad (4\cdot26)$$

i.e.

$$\sum_{k}^{N}\frac{m_p^2}{\tau+\dfrac{\nu}{\alpha_p}}=\frac{1}{1+\tau}. \quad (4\cdot27)$$

We see that the effect of the regression variables chosen as latent vectors has been the same as on previous occasions, i.e. the corresponding latent roots are eliminated from the problem. We may therefore without loss of generality suppose that

$$\alpha_k<\alpha_{k+1}<\dots<\alpha_N. \quad (4\cdot28)$$

The location of the roots ν_i is easily seen graphically from (4·27). We may, however, proceed as follows. Define the continuous function of ν,

$$f(\nu)=\prod_{p=k}^{N}\left(\tau+\frac{\nu}{\alpha_p}\right)-(1+\tau)\sum_{p=k}^{N}m_p^2\prod_{j\neq p}^{N}\left(\tau+\frac{\nu}{\alpha_j}\right).$$

Then since

$$f(-\tau\alpha_i)=-(1+\tau)m_i^2\tau^{N-k}\prod_{\substack{j=k\\ j\neq i}}^{N}\left(1-\frac{\alpha_i}{\alpha_j}\right)$$

is seen to have the opposite sign to $f(-\tau\alpha_{i+1})$, the equation (4·26) must have at least one root in each of the intervals

$$(-\tau\alpha_{i+1}, -\tau\alpha_i).$$

But, considering equation (4·27), we see that

$$\sum_{k}^{N}\frac{m_i^2}{\tau+\dfrac{\alpha_N}{\alpha_i}}\leqslant\frac{1}{1+\tau}, \quad \sum_{k}^{N}\frac{m_i^2}{\tau+\dfrac{\alpha_k}{\alpha_i}}\geqslant\frac{1}{1+\tau}$$

by (4·28) and the fact that $\sum_{k}^{N}m_i^2=1$. Hence there is at least one root of (4·26) and (4·27) in the interval

$$(\alpha_k, \alpha_N).$$

Combining the results, we have proved that there is one root ν_i in each of the $N-k+1$ intervals

$$(-\tau\alpha_{i+1}, -\tau\alpha_i) \quad (i=k,\dots,N-1), \qquad (\alpha_k, \alpha_N). \quad (4\cdot29)$$

Finally, if $\mathbf{x}_k$ is chosen to be the latent vector associated with the root α_k, i.e. $\mathbf{a}_k$, (4·26) gives us the roots

$$-\tau\alpha_N, \ldots, -\tau\alpha_{k+1} \quad \text{and} \quad \alpha_k, \tag{4·30}$$

which are seen from (4·29) to be the least possible set of roots. If $\mathbf{x}_k$ is chosen to be the latent vector associated with α_N, i.e. $\mathbf{a}_N$, (4·26) gives the greatest possible set of roots

$$-\tau\alpha_{N-1}, \ldots, -\tau\alpha_k \quad \text{and} \quad \alpha_N. \tag{4·31}$$

We have therefore proved

THEOREM 1. If τ is such that $P(T<\tau) = C\ (0 \leqslant C \leqslant 1)$, T being defined by (4·20), then there exist numbers τ_L, τ_U such that

$$\tau_L \leqslant \tau \leqslant \tau_U \tag{4·32}$$

for all vectors $\mathbf{x}_k$. τ_L and τ_U are determined by the equations

$$\left.\begin{aligned} P\left(\frac{\alpha_k\eta_k^2}{\alpha_{k+1}\eta_{k+1}^2+\ldots+\alpha_N\eta_N^2} < \tau_L\right) &= C, \\ P\left(\frac{\alpha_N\eta_N^2}{\alpha_k\eta_k^2+\ldots+\alpha_{N-1}\eta_{N-1}^2} < \tau_U\right) &= C, \end{aligned}\right\} \tag{4·33}$$

where $\eta_k, \ldots, \eta_{N-1}$ are N.I.D. $(0, 1)$.

We now extend this theorem, with the same assumptions about the model, to a joint test of the null hypothesis

$$\beta_1 = \beta_1^0, \quad \ldots \quad \beta_h = \beta_h^0 \quad (h \leqslant k-1), \qquad \beta_k = \beta_k^0. \tag{4·34}$$

The statistic T for this hypothesis is on the null hypothesis

$$T = \frac{\mathbf{u}'\mathbf{a}_1\mathbf{a}_1'\mathbf{u}+\ldots+\mathbf{u}'\mathbf{a}_h\mathbf{a}_h'\mathbf{u}+\mathbf{u}'\mathbf{x}_k\mathbf{x}_k'\mathbf{u}}{\mathbf{u}'(\mathbf{I}-\mathbf{a}_1\mathbf{a}_1'-\ldots-\mathbf{a}_{k-1}\mathbf{a}_{k-1}'-\mathbf{x}_k\mathbf{x}_k')\mathbf{u}}. \tag{4·35}$$

Proceeding as before the determinantal equation is found to be

$$|\,\mathbf{a}_1\mathbf{a}_1'+\ldots+\mathbf{a}_h\mathbf{a}_h'+\mathbf{x}_k\mathbf{x}_k'-\tau(\mathbf{I}-\mathbf{a}_1\mathbf{a}_1'-\ldots-\mathbf{a}_{k-1}\mathbf{a}_{k-1}'-\mathbf{x}_k\mathbf{x}_k')-\nu\boldsymbol{\alpha}^{-1}\,| = 0. \tag{4·36}$$

After applying the orthogonal transformation K of (4·23) the roots ν_i are seen to be

$$\alpha_1, \alpha_2, \ldots, \alpha_h, \quad k-h-1 \text{ zeros},$$

and the $N-k+1$ roots of (4·25). This leads immediately to the following extension of Theorem 1.

THEOREM 2. If τ is such that $P(T<\tau) = C, (0 \leqslant C \leqslant 1)$ T being defined by (4·35), then there exist numbers τ_L and τ_U such that

$$\tau_L \leqslant \tau \leqslant \tau_U$$

for all vectors x_k. τ_L and τ_U are determined by the equations

$$\left.\begin{aligned} P\left(\frac{\alpha_1\eta_1^2+\ldots+\alpha_h\eta_h^2+\alpha_k\eta_k^2}{\alpha_{h+1}\eta_{k+1}^2+\ldots+\alpha_N\eta_N^2} < \tau_L\right) &= C, \\ P\left(\frac{\alpha_1\eta_1^2+\ldots+\alpha_h\eta_h^2+\alpha_N\eta_N^2}{\alpha_k\eta_k^2+\ldots+\alpha_{N-1}\eta_{N-1}^2} < \tau_U\right) &= C, \end{aligned}\right\} \tag{4·37}$$

where $\eta_1, \ldots, \eta_h, \eta_k, \ldots, \eta_N$ are N.I.D. $(0, 1)$.

To complete this case, we examine the test of the null hypothesis

$$\beta_1 = \beta_1^0, \ldots, \beta_h = \beta_h^0 \quad (h \leqslant k-1). \tag{4·38}$$

The T statistic is here defined on the null hypothesis by

$$T = \frac{\mathbf{u}'\mathbf{a}_1\mathbf{a}_1'\mathbf{u} + \ldots + \mathbf{u}'\mathbf{a}_h\mathbf{a}_h'\mathbf{u}}{\mathbf{u}'(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1' - \ldots - \mathbf{a}_{k-1}\mathbf{a}_{k-1}' - \mathbf{x}_k\mathbf{x}_k')\mathbf{u}}. \tag{4·39}$$

The associated determinantal equation is

$$|\mathbf{a}_1\mathbf{a}_1' + \ldots + \mathbf{a}_h\mathbf{a}_h' - \tau(\mathbf{I} - \mathbf{a}_1\mathbf{a}_1' - \ldots - \mathbf{a}_{k-1}\mathbf{a}_{k-1}' - \mathbf{x}_k\mathbf{x}_k') - \nu\boldsymbol{\alpha}^{-1}| = 0. \tag{4·40}$$

The roots of (4·40) are $\alpha_1, \ldots, \alpha_h$, $k-h$ zeros and the roots of the determinantal equation

$$\begin{vmatrix} \tau m_k^2 - \tau - \dfrac{\nu}{\alpha_k}, & \tau m_k m_{k+1}, & \ldots, & \tau m_k m_N \\ \tau m_{k+1} m_k, & \tau m_{k+1}^2 - \tau - \dfrac{\nu}{\alpha_{k+1}}, & \ldots, & \tau m_{k+1} m_N \\ \vdots & \vdots & & \vdots \\ \tau m_N m_k, & \tau m_N m_{k+1}, & \ldots, & \tau m_N^2 - \tau - \dfrac{\nu}{\alpha_N} \end{vmatrix} = 0. \tag{4·41}$$

Expanding as before, we find the equation in ν

$$\prod_{i=k}^{N}\left(\tau + \frac{\nu}{\alpha_i}\right) - \tau \sum_{i=k}^{N} m_i^2 \prod_{j \neq i}\left(\tau + \frac{\nu}{\alpha_j}\right) = 0, \tag{4·42}$$

i.e.

$$\frac{1}{\tau} = \sum_{i=k}^{N} \frac{m_i^2}{\tau + \dfrac{\nu}{\alpha_i}}. \tag{4·43}$$

As with equation (4·26), it can be shown that equation (4·42) has at least one root in each of the intervals

$$(-\tau\alpha_{i+1}, -\tau\alpha_i).$$

By inspection we see that it always has a zero root so that there is just one in each of these intervals. If $\mathbf{x}_k$ is chosen to be the latent vector of $\boldsymbol{\alpha}$ associated with the root α_k, i.e. $\mathbf{a}_k$, the least set of roots for ν_i is obtained, while if $\mathbf{x}_k$ is the latent vector of $\boldsymbol{\alpha}$ associated with the root α_N, i.e. $\mathbf{a}_N$, the greatest set is found. We may thus state

THEOREM 3. If τ is such that $P(T < \tau) = C$ $(0 \leqslant C \leqslant 1)$, T being defined by (4·39), then there exist numbers τ_L and τ_U such that

$$\tau_L \leqslant \tau \leqslant \tau_U$$

for all vectors x_k. τ_L and τ_U are determined by the equations

$$\left.\begin{aligned} P\left(\frac{\alpha_1\eta_1^2 + \ldots + \alpha_h\eta_h^2}{\alpha_{k+1}\eta_{k+1}^2 + \ldots + \alpha_N\eta_N^2} < \tau_L\right) &= C, \\ P\left(\frac{\alpha_1\eta_1^2 + \ldots + \alpha_h\eta_h^2}{\alpha_k\eta_k^2 + \ldots + \alpha_{N-1}\eta_{N-1}^2} < \tau_U\right) &= C, \end{aligned}\right\} \tag{4·44}$$

where $\eta_1 \ldots \eta_h, \eta_k \ldots \eta_N$ are N.I.D. (0, 1).

We now proceed to a case where there is more than one arbitrary regression vector. Suppose that all k regression vectors are arbitrary except for restriction (2·2) and that the hypothesis to be tested is

$$\beta_i = \beta_i^0 \quad (i = 1, \ldots, k). \tag{4·45}$$

On the null hypothesis the usual F-ratio is distributed like a multiple of

$$T = \frac{\mathbf{u}'\mathbf{X}\mathbf{X}'\mathbf{u}}{\mathbf{u}'(\mathbf{I}-\mathbf{X}\mathbf{X}')\mathbf{u}}; \tag{4·46}$$

therefore

$$R = \frac{T}{1+T} = \frac{\mathbf{u}'\mathbf{X}\mathbf{X}'\mathbf{u}}{\mathbf{u}'\mathbf{u}}. \tag{4·47}$$

Since T is a monotonic function of $R\left(T = \frac{R}{1-R}\right)$, the extremes of $P(R<r)$ will yield extremes of $P(T<\tau)$. We therefore proceed to find distributions which bound that of R above and below. If $\mathbf{H}$ is the linear transformation such that

$$\mathbf{H}\alpha\mathbf{H}' = \mathbf{I}, \tag{4·48}$$

then the variables $\mathbf{v} = \mathbf{H}\mathbf{u}$ are N.I.D. $(0, 1)$—for σ^2 may be set equal to unity without loss of generality. Then

$$R = \frac{\mathbf{v}'\mathbf{H}'^{-1}\mathbf{X}\mathbf{X}'\mathbf{H}^{-1}\mathbf{v}}{\mathbf{v}'(\mathbf{H}\mathbf{H}')^{-1}\mathbf{v}}.$$

From (4·48),

$$(\mathbf{H}\mathbf{H}')^{-1} = \begin{bmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_N \end{bmatrix},$$

so that we may write

$$R = \frac{\mathbf{v}'\mathbf{H}'^{-1}\mathbf{X}\mathbf{X}'\mathbf{H}^{-1}\mathbf{v}}{\sum_1^N \alpha_i v_i^2}. \tag{4·49}$$

R is a ratio of non-negative definite quadratic forms in N.I.D. $(0, 1)$ variables. For variation of $\mathbf{X}$, the upper and lower bounding random variables for R are obtained when the numerator takes its least and greatest forms. Now the numerator of R is distributed like

$$\sum_{i=1}^{k} \nu_i \eta_i^2, \tag{4·50}$$

where $\eta_1, \ldots, \eta_k$ are N.I.D. $(0, 1)$ and $\nu_1, \ldots, \nu_k$ are the k non-zero roots, written in ascending order of magnitude, of

$$|\mathbf{H}'^{-1}\mathbf{X}\mathbf{X}'\mathbf{H}^{-1} - \nu\mathbf{I}| = 0, \tag{4·51}$$

i.e. the non-zero roots of

$$|\mathbf{H}^{-1}\mathbf{H}'^{-1}\mathbf{X}\mathbf{X}' - \nu\mathbf{I}| = 0,$$

since the roots of a product are indifferent to the order of multiplication. But by (4·48)

$$\mathbf{H}^{-1}\mathbf{H}'^{-1} = (\mathbf{H}'\mathbf{H})^{-1} = \alpha,$$

so that $\nu_1 < \ldots < \nu_k$ are the k non-zero roots of

$$|\alpha\mathbf{X}\mathbf{X}' - \nu\mathbf{I}| = 0. \tag{4·52}$$

Suppose that $\mathbf{z}_1, \ldots, \mathbf{z}_{N-k}$ are vectors which together with $\mathbf{x}_1, \ldots, \mathbf{x}_k$ form a complete orthonormal set in N-space. If $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_{N-k}]$, then,

$$[\mathbf{X}, \mathbf{Z}][\mathbf{X}, \mathbf{Z}]' = [\mathbf{X}, \mathbf{Z}]'[\mathbf{X}, \mathbf{Z}]$$
$$= \mathbf{I}.$$

But $$[\mathbf{X}, \mathbf{Z}][\mathbf{X}, \mathbf{Z}]' = \mathbf{XX}' + \mathbf{ZZ}',$$

so that
$$\begin{aligned}\mathbf{XX}' &= \mathbf{I} - \mathbf{ZZ}' \\ &= (\mathbf{I} - \mathbf{z}_1\mathbf{z}_1') \dots (\mathbf{I} - \mathbf{z}_{N-k}\mathbf{z}_{N-k}') \\ &= \mathbf{M}_1 \dots \mathbf{M}_{N-k}, \quad \text{say.}\end{aligned}$$

where $\mathbf{M}_1, \dots, \mathbf{M}_{N-k}$ are idempotent matrices of rank $N-1$.

With this characterization of $\mathbf{XX}'$, inequalities on the roots of (4·52) may be obtained immediately from the investigation of Durbin & Watson (1950). They show, with a proof similar to that of Theorem 1, that the roots of $\boldsymbol{\alpha}\mathbf{M}_1$ lie between the roots of $\boldsymbol{\alpha}$. Thus the non-zero roots of $\boldsymbol{\alpha}\mathbf{M}_1\mathbf{M}_2$ lie between the non-zero roots of $\boldsymbol{\alpha}\mathbf{M}_1$. Combining the inequalities given by successive applications of this result, it follows that the roots $\nu_1, \dots, \nu_k$ of (4·52) must satisfy the inequalities

$$\alpha_1 \leqslant \nu_1 \leqslant \alpha_{N-k+1}, \quad \alpha_2 \leqslant \nu_2 \leqslant \alpha_{N-k+2}, \quad \dots, \quad \alpha_k \leqslant \nu_k \leqslant \alpha_N. \tag{4·53}$$

The least set of roots is obtained when the k regression vectors are chosen to be the latent vectors $\mathbf{a}_1, \dots, \mathbf{a}_k$ (or linear combinations of these vectors). The greatest set of roots is obtained when the regression vectors are similarly related to the latent vectors $\mathbf{a}_{N-k+1}, \dots, \mathbf{a}_N$. These choices lead to the extreme forms of R required. We may thus state

THEOREM 4. If τ is such that $P(T < \tau) = C$ $(0 \leqslant C \leqslant 1)$, T being defined by (4·46). then there exist numbers τ_L and τ_U such that

$$\tau_L \leqslant \tau \leqslant \tau_U$$

for all $\mathbf{X}$. τ_L and τ_U are defined by the equations

$$\left.\begin{aligned} P\left(\frac{\alpha_1\eta_1^2 + \dots + \alpha_k\eta_k^2}{\alpha_{k+1}\eta_{k+1}^2 + \dots + \alpha_N\eta_N^2} < \tau_L\right) &= C, \\ P\left(\frac{\alpha_{N-k+1}\eta_{N-k+1}^2 + \dots + \alpha_N\eta_N^2}{\alpha_1\eta_1^2 + \dots + \alpha_{N-k}\eta_{N-k}^2} < \tau_U\right) &= C, \end{aligned}\right\} \tag{4·54}$$

where $\eta_1, \dots, \eta_N$ are N.I.D. $(0, 1)$.

The rules of formation of the random variables which lead to the bounding significance points are now evident from Theorems 1, 2, 3, 4 so that there is no need to give the details for every type of test which may arise. In all cases it is seen that the distributions of the bounding variables converge to the appropriate F-distribution as the α's approach equality.

The calculation of the bounding significance points discussed above presents some difficulty. We are faced with the determination of the significance points of ratio of two independent weighted sums of squares of normal deviates. Since the weights are all positive. the method of Pitman & Robbins (1949) may be used when the α's are not too different. Some numerical results obtained by this method will be given in Part II. The amount of arithmetic required is always large—the electronic computer of the Cambridge Mathematical Laboratory was used in these calculations—so it is worth while examining approximate methods.

To illustrate the approximations used in Part II, let us consider the random variable which generates the lower bound in Theorem 4, namely,

$$\frac{\alpha_1\eta_1^2 + \dots + \alpha_k\eta_k^2}{\alpha_{k+1}\eta_{k+1}^2 + \dots + \alpha_N\eta_N^2}.$$

The simplest approximation to this ratio is

$$\frac{\sum_1^k \alpha_i}{k} \frac{N-k}{\sum_{k+1}^N \alpha_i} \frac{\eta_1^2+\ldots+\eta_k^2}{\eta_{k+1}^2+\ldots+\eta_N^2}, \quad \text{i.e.} \quad \frac{\sum_1^k \alpha_i}{\sum_{k+1}^N \alpha_i} F_{k,N-k},$$

where $F_{k,N-k}$ has the F-distribution with k and $N-k$ degrees of freedom. If in a practical analysis we had some knowledge of the correlogram of the errors, this approach suggests that multiples of the usual significance points might be used to make a reasonable test. A better approximation may be obtained by replacing the numerator and denominator by multiples of σ^2 with the correct variances as well as the correct means. The significance points of this approximation must be obtained by interpolation and do not have the simple interpretation of the first approximation.

The author has pleasure in recording his indebtedness to Mr J. Durbin for many helpful discussions.

APPENDIX

A proof of the inequality (3·2)

THEOREM. Let $a_j > 0$, $b_j > 0$, $w_j \geqslant 0$ $(j = 1, \ldots, n)$. Then

$$1 \leqslant \frac{(\Sigma a_j^2 w_j)(\Sigma b_j^2 w_j)}{(\Sigma a_j b_j w_j)^2} \leqslant \max_{i,j} \frac{(a_i b_j + a_j b_i)^2}{4 a_i a_j b_i b_j}.$$

The extreme values are attained where at most two of $w_1, \ldots, w_n$ are non-zero.

This theorem and its proof below are due to J. W. S. Cassels.

Proof. As a straightforward extremal problem it may be shown that

$$1 \leqslant \frac{(1+kw)(1+k^{-1}w)}{(1+w)^2} \leqslant \frac{(1+k)(1+k^{-1})}{4} \quad \text{if} \quad k > 0,\ w \geqslant 0.$$

Thus the theorem holds for $n = 2$—for put

$$w = \frac{a_2 b_2 w_2}{a_1 b_1 w_1}, \quad k = \frac{a_1 b_2}{a_2 b_1}.$$

Furthermore, the extremes of the theorem are attained when at most two of $w_1, \ldots, w_n$ are non-zero. It may be assumed without loss of generality that

$$\begin{vmatrix} a_i^2 & a_j^2 & a_k^2 \\ b_i & b_j & b_k \\ a_i b_i & a_j b_j & a_k b_k \end{vmatrix} \neq 0.$$

For this determinant is given by $\prod_{i,j,k} \begin{vmatrix} a_i & a_j \\ b_i & b_j \end{vmatrix}$, so that if it vanishes we have say

$$a_i = \lambda a_j, \quad b_i = \lambda b_j, \quad \text{where} \quad \lambda > 0.$$

But if this is so the problem is reduced to one in $n-1$ variables.

To prove that the extrema of $(\Sigma a_j^2 w_j)(\Sigma b_j^2 w_j)/(\Sigma a_j b_j w_j)^2$ are attained when at most two of $w_1 \dots w_N$ are non-zero, let M be an extreme which is attained when say $w_1 \neq 0, w_2 \neq 0, w_3 \neq 0$. Then we have

$$\frac{d}{dw_k}\{(\Sigma a_j^2 w)(\Sigma b_j^2 w_j) - M(\Sigma a_j b_j w_j)^2\} = 0 \quad (k = 1, 2, 3),$$

and so

$$a_k^2 X + b_k^2 Y - 2Ma_k b_k Z = 0 \quad (k = 1, 2, 3),$$

where $X = \Sigma a_j^2 w_j$, $Y = \Sigma b_j^2 w_j$, $Z = \Sigma a_j b_j w_j$. But the determinant of these equations does not vanish. Hence

$$X = Y = MZ = 0,$$

which is impossible.

This establishes the theorem. To obtain the inequality (3·2), it is only necessary to notice that the upper bound in theorem may be put in the form

$$\max_{i,j} \frac{1}{4}\left(\frac{r_i}{r_j} + \frac{r_j}{r_i}\right)^2,$$

where $r_i = a_i/b_i$. Since the function $w + 1/w$ takes its maximum at the end of the range of variation of w, (3·2) follows.

REFERENCES

AITKEN, A. C. (1934–5). On least squares and linear combinations of observations. *Proc. Roy. Soc. Edinb.* **55**, 42–8.

AITKEN, A. C. (1948). On the estimation of many statistical parameters. *Proc. Roy. Soc. Edinb.* **62**, 369.

ANDERSON, R. L. & ANDERSON, T. W. (1950). Distribution of the circular serial correlation coefficient for residuals from a fitted Fourier series. *Ann. Math. Statist.* **21**, 59–81.

BOX, G. E. P. (1954). Some theorems on quadratic forms applied to the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Ann. Math. Statist.* **25**, 484–98.

COCHRANE, D. & ORCUTT, G. H. (1949). Application of least squares regression to relationships containing autocorrelated error terms. *J. Amer. Statist. Ass.* **44**, 32–61.

DURBIN, J. & WATSON, G. S. (1950). Testing for serial correlation in least squares regression. I. *Biometrika*, **37**, 409–28.

DURBIN, J. & WATSON, G. S. (1951). Testing for serial correlation in least squares regression. II. *Biometrika*, **38**, 159–78.

HANNAN, E. J. (1955). Exact tests for serial correlation. *Biometrika*, **42**, 133.

HARDY, G. H., LITTLEWOOD, J. E. & POLYA, G. (1934). *Inequalities*. Cambridge University Press.

PITMAN, E. J. G. & ROBBINS, H. (1949). Application of the method of mixtures to quadratic forms in normal variates. *Ann. Math. Statist.* **20**, 552–60.

WOLD, H. (1949). On least squares regression with autocorrelated variables and residuals. *Trans. Int. Inst. Statist.* Reprint, pp. 1–13.

SOME THEOREMS AND SUFFICIENCY CONDITIONS FOR THE MAXIMUM-LIKELIHOOD ESTIMATOR OF AN UNKNOWN PARAMETER IN A SIMPLE MARKOV CHAIN

By J. GANI

Australian National University, Canberra, A.C.T.

I. Summary

The paper begins with proofs of the usual theorems for the optimum properties of the maximum-likelihood estimator of an unknown parameter θ which defines the transition probabilities $p_{ij}(\theta)$ of a simple ergodic Markov chain. By an ergodic chain is meant one for which, not only is the final chain stationary, but also all possible initial states remain permanently available; these conditions are sufficient to prove that the maximum-likelihood estimator is consistent, and asymptotically normally distributed.

The paper proceeds to establish the form of the transition probabilities $p_{ij}(\theta)$ which admit a sufficient estimator of θ. To do this, the form of the likelihood function admitting a sufficient estimator when the parent distribution is discrete is first derived: this is used to obtain the form of the probabilities $p_i(\theta)$ for a multinomial distribution admitting a sufficient estimator of θ, and the result is finally generalized for the transition probabilities $p_{ij}(\theta)$ of the simple ergodic Markov chain.

The paper closes with an examination of possible forms for the matrix $\mathbf{p}$ of transition probabilities $p_{ij}(\theta)$, and these are illustrated with simple examples for Markov chains with two and three states.

II. The likelihood equation for a simple Markov chain

Consider the simple Markov chain with s possible states $E_1, \ldots, E_s$, and the matrix of transition probabilities $p_{ij}(\theta) = \Pr(E_j \mid E_i)\,(i,j = 1, \ldots, s)$, which are all functions of an unknown parameter θ. The matrix is given by

$$\mathbf{p}(\theta) = \begin{pmatrix} p_{11}(\theta) & \ldots & p_{1s}(\theta) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ p_{s1}(\theta) & \ldots & p_{ss}(\theta) \end{pmatrix},$$

where the transition probabilities are subject to the condition that

$$\sum_{j=1}^{s} p_{ij} = 1, \quad \text{for all } i = 1, 2, \ldots, s. \tag{1}$$

If a realization of the chain results in an observed sequence S of $n+1$ states in the following order

$$E_i, E_j, \ldots, E_k, E_l,$$

then it is possible to write the probabilities of this sequence as

$$\phi(S) = a_i(\theta)\, p_{ij}(\theta) \ldots p_{kl}(\theta), \tag{2}$$

where $a_i(\theta)$ $(i = 1, \ldots, s)$, the initial probability distribution of the states $E_1, \ldots, E_s$, is assumed known.

The likelihood function will be given by

$$L = \ln \phi(S) = \ln a_i + \ln p_{ij} + \ldots + \ln p_{kl}.$$

Grouping together the transitions from states E_i to states E_j, $(i,j = 1, \ldots, s)$, and denoting the frequency of these transitions by n_{ij}, we write this likelihood function as

$$L = \ln a_i + \sum_{i=1}^{s} \sum_{j=1}^{s} n_{ij} \ln p_{ij},$$

or for simplicity

$$L = \ln a_i + \sum_{i,j} n_{ij} \ln p_{ij}. \tag{3}$$

It follows that, providing $a_i(\theta)$, $p_{ij}(\theta)$ are differentiable with respect to θ, the derivative of the likelihood is

$$\frac{dL}{d\theta} = \frac{1}{a_i}\frac{da_i}{d\theta} + \sum_{i,j} \frac{n_{ij}}{p_{ij}} \frac{dp_{ij}}{d\theta}. \tag{4}$$

As n increases, the second parts of (3) and (4) become dominant, and the likelihood function and its derivative will be

$$L \sim \sum_{i,j} n_{ij} \ln p_{ij}, \tag{5}$$

and

$$\frac{dL}{d\theta} \sim \sum_{i,j} \frac{n_{ij}}{p_{ij}} \frac{dp_{ij}}{d\theta}. \tag{6}$$

The maximum-likelihood estimator T of θ will in this case be given by

$$\left(\frac{dL}{d\theta}\right)_T \sim \left(\sum_{i,j} \frac{n_{ij}}{p_{ij}} \frac{dp_{ij}}{d\theta}\right)_T = 0. \tag{7}$$

It is important to note that the n_{ij} are not linearly independent; the number of transitions from the state E_i to the states $E_1, \ldots, E_s$, will, except for a possible end-effect, be equal to the number of transitions from the states $E_1, \ldots, E_s$, to the state E_i, so that

$$\sum_{j=1}^{s} n_{ij} \doteq \sum_{r=1}^{s} n_{ri}, \tag{8}$$

where the sign $\doteq$ indicates equality or a possible difference of 1 between the sums. As n increases, we may therefore accept the equations

$$\sum_{j=1}^{s} n_{ij} = \sum_{r=1}^{s} n_{ri} = n_i \quad (i = 1, \ldots, s), \tag{9}$$

where

$$\sum_{i=1}^{s} n_i = \sum_{i=1}^{s} \sum_{j=1}^{s} n_{ij} = n.$$

III. Optimum properties of the maximum-likelihood estimator

Following closely the proofs given by Cramér (1946) and Rao (1952) for continuous distributions, we shall deduce that for an ergodic chain, the maximum-likelihood estimator T obtained as a solution of (7) has the following optimum properties:

(*a*) The estimator T obtained as a solution of (7) is consistent.

(*b*) The consistent solution of (7) is asymptotically normally distributed about the true value θ_0: it is fully efficient.

A further optimum property due to Huzurbazar (1948) can also be proved; it is that

(*c*) Any consistent solution of (7) is such that for $n \to \infty$ the probability that the likelihood tend to a maximum converges to 1; the consistent solution is therefore unique.

The proof of this property for the case of the chain considered is, however, identical in its essentials with that given by Huzurbazar for continuous distributions; it will therefore be omitted. In order to establish these properties, we require the assumptions that

$$\frac{dp_{ij}}{d\theta}, \quad \frac{d^2p_{ij}}{d\theta^2}, \quad \frac{d^3p_{ij}}{d\theta^3}$$

exist and are continuous for every θ in a range R including the true value θ_0. Before proceeding to prove these theorems, it may be useful to mention some properties of ergodic chains which we shall refer to.

For an ergodic chain, all possible initial states remain permanently available, and the final chain is stationary and independent of initial conditions. Although it is possible for some transition probabilities p_{ij} of an ergodic chain to be zero, the stationary probabilities $P_i = \Pr(E_i)$ $(i = 1, \ldots, s)$, which are given by the matrix equation

$$\mathbf{p}'\mathbf{P} = \mathbf{P},$$

together with the equation $\sum_i P_i = 1$, where $\mathbf{P}$ is the column vector of elements P_i, will all be non-zero; it is clear that the P_i will also be functions of the parameter θ. In equation (2) no assumption was made about the chain being initially stationary, so that the a_i are not necessarily equal to the P_i; however, if the process is initially stationary, we can write $a_i = P_i$ in this and the subsequent equations (3) and (4). In all that follows, we shall for simplicity consider an initially stationary chain; the results obtained will also hold asymptotically for a chain which is not initially stationary.

The following set of results for an ergodic chain is connected with the matrix $\mathbf{R}$, with a transpose of form

$$\mathbf{R}' = \begin{pmatrix} p_{11}e^{t_{11}} & p_{12}e^{t_{12}} & \ldots & p_{1s}e^{t_{1s}} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ p_{s1}e^{t_{s1}} & \ldots & \ldots & p_{ss}e^{t_{ss}} \end{pmatrix}, \tag{10}$$

which was introduced by Montroll (1947) and Bartlett (1951), and will later be used to obtain the moment generating function of the transition frequencies n_{ij}, previously mentioned in equation (5). It is easy to see that the latent roots $\mu_r(\mathbf{t})$ of this matrix, where $\mathbf{t}$ is the matrix of elements t_{ij} $(i, j = 1, \ldots, s)$ are given by the equation

$$|\mathbf{R} - \mu\mathbf{I}| = 0 \tag{11}$$

and are continuous in $\mathbf{t}$. For $\mathbf{t} = 0$, this determinantal equation becomes

$$|\mathbf{p} - \mu\mathbf{I}| = 0, \tag{12}$$

with roots $\mu_1(0), \ldots, \mu_s(0)$, not necessarily all distinct; we may, without loss of generality, assume that these are arranged so that their moduli are in descending order of magnitude. For a stationary chain, it is known that

$$\mu_1(0) = 1, \quad |\mu_r(0)| < 1 \quad \text{for the remaining } r = 2, 3, \ldots, s;$$

it follows from the continuity of the roots $\mu_1(\mathbf{t}), \ldots, \mu_s(\mathbf{t})$, that for $\mathbf{t}$ in the neighbourhood of $\mathbf{t} = 0$, we have

$$|\mu_r(\mathbf{t})| < |\mu_1(\mathbf{t})| \quad \text{for all } r = 2, 3, \ldots, s.$$

We prove also that for an ergodic chain, for some $\mathbf{t}$ in the neighbourhood of $\mathbf{t} = 0$, $\mu_1(\mathbf{t})$ is not identically equal to 1. For, suppose $\mu_1(\mathbf{t}) \equiv 1$, then for $\mathbf{t}$ such that $t_{11} \neq 0$, and $t_{ij} = 0$ for all other values of i, j, the equation (11) would give

$$\begin{vmatrix} p_{11}e^{t_{11}} - 1 & p_{12} & \dots & p_{1s} \\ p_{21} & p_{22} - 1 & \dots & p_{2s} \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ p_{s1} & \dots & \dots & p_{ss} - 1 \end{vmatrix} = 0. \tag{13}$$

On expansion, this could be written

$$(p_{11}e^{t_{11}} - 1)C_{11} + p_{12}C_{12} + \dots + p_{1s}C_{1s} = 0,$$

where the C_{1j} are cofactors of the elements in the first row and jth column. For t_{11} small, this is

$$p_{11}t_{11}C_{11} + (p_{11} - 1)C_{11} + p_{12}C_{12} + \dots + p_{1s}C_{1s} = 0.$$

Now if $t_{11} = 0$ also, so that $\mathbf{t} = 0$, equation (13) would give

$$(p_{11} - 1)C_{11} + p_{12}C_{12} + \dots + p_{1s}C_{1s} = 0;$$

we see, therefore, that $\mu_1(\mathbf{t}) \equiv 1$ only if $p_{11}C_{11}t_{11} = 0$, so that $p_{11} = 0$ or $C_{11} = 0$, or both are zero. C_{11} cannot be zero, for since $\mu_1(0) = 1$ is a simple root, then

$$\left\{\frac{d}{d\mu}(|\mathbf{p} - \mu\mathbf{I}|)\right\}_{\mu=1} \neq 0,$$

or, on expansion,

$$-\{C_{11} + C_{22} + \dots + C_{ss}\} \neq 0,$$

where C_{ii} are cofactors of the elements in the leading diagonal of (13) when $t_{11} = 0$. At least one of these C_{ii} is non-zero, and we may without loss of generality assume that C_{11} is such a non-zero cofactor. If, in addition, p_{11} is non-zero, $\mu_1(\mathbf{t}) \neq 1$ for at least the case when $t_{11} \neq 0$ and $t_{ij} = 0$ for all other i, j.

It is possible, however, that p_{11} be zero; if this is so, then for an ergodic chain, at least two of the probabilities $p_{12}, \dots, p_{1s}$ will be non-zero.* From equation (12) we see that in this case

$$-C_{11} + p_{12}C_{12} + \dots + p_{1s}C_{1s} = 0,$$

so that it is clear that at least one of the cofactors C_{1r} $(r = 2, \dots, s)$ which multiplies a non-zero probability p_{1r} is non-zero. Let $\mathbf{t}$ now be such that $t_{1r} \neq 0$, and the remaining t_{ij} are all zero; then in exactly the same way as before, for a small value of t_{1r}, it can be shown that $\mu_1(\mathbf{t}) \equiv 1$ only if $p_{1r}C_{1r}t_{1r} = 0$, a condition which is absurd. So that for $p_{11} = 0$, there still exists a $\mathbf{t}$ for which $\mu_1(\mathbf{t}) \neq 1$. It follows therefore that in general for an ergodic chain, for some $\mathbf{t}$ in the neighbourhood of $\mathbf{t} = 0$, $\mu_1(\mathbf{t})$ is not identically equal to 1.

(1) *Consistency of the maximum-likelihood estimator*

If the estimator T obtained as a solution of (7) is to be consistent, it is necessary that if θ_0 is the real value of θ, then as $n \to \infty$,

$$\Pr\{|T - \theta_0| < \delta\} \to 1, \tag{14}$$

where δ is any small positive number.

* By an ergodic chain is usually understood one for which $0 \leqslant p_{ij} \leqslant 1$, but no $p_{ii} = 1$ $(i, j = 1, \dots, s)$. It is possible, however, that a particular $p_{ij} = 1$ for $i \neq j$; in this case the states E_i, E_j may be contracted, and we may therefore accept that for an ergodic chain, $0 \leqslant p_{ij} < 1$ for all $i, j = 1, \dots, s$. Our statement follows immediately.

Consider the expansion

$$\frac{dL}{d\theta} = \left\{\frac{dL}{d\theta}\right\}_{\theta_0} + (\theta-\theta_0)\left\{\frac{d^2L}{d\theta^2}\right\}_{\theta_0} + \tfrac{1}{2}(\theta-\theta_0)^2\left\{\frac{d^3L}{d\theta^3}\right\}_{\theta_1}, \tag{15}$$

where θ_1 lies in (θ, θ_0). For simplicity, we write $\left\{\frac{dL}{d\theta}\right\}_{\theta_0}$, $\left\{\frac{d^2L}{d\theta^2}\right\}_{\theta_0}$, $\left\{\frac{d^3L}{d\theta^3}\right\}_{\theta_1}$ as nB_0, nB_1, nB_2 respectively, where these are obtained from (6) as

$$\left.\begin{aligned} nB_0 &= \left\{\sum_{i,j}\frac{n_{ij}}{p_{ij}}\frac{dp_{ij}}{d\theta}\right\}_{\theta_0}, \\ nB_1 &= \left\{\sum_{i,j} n_{ij}\left[\frac{1}{p_{ij}}\frac{d^2p_{ij}}{d\theta^2} - \frac{1}{p_{ij}^2}\left(\frac{dp_{ij}}{d\theta}\right)^2\right]\right\}_{\theta_0}, \\ nB_2 &= \left\{\sum_{i,j} n_{ij}\left[\frac{1}{p_{ij}}\frac{d^3p_{ij}}{d\theta^3} - \frac{3}{p_{ij}^2}\frac{dp_{ij}}{d\theta}\frac{d^2p_{ij}}{d\theta^2} + \frac{2}{p_{ij}^3}\left(\frac{dp_{ij}}{d\theta}\right)^3\right]\right\}_{\theta_1}. \end{aligned}\right\} \tag{16}$$

It is simple to see that as $n\to\infty$, the expectations of the n_{ij}, irrespective of whether the chain is initially or only finally stationary, will be given for all $i,j = 1, 2, \ldots, s$, by

$$\mathscr{E}(n_{ij}) = nP_ip_{ij}.$$

Further, since on differentiation with respect to θ, equation (1) gives

$$\sum_{j=1}^{s}\frac{dp_{ij}}{d\theta} = 0, \quad \sum_{j=1}^{s}\frac{d^2p_{ij}}{d\theta^2} = 0, \quad \sum_{j=1}^{s}\frac{d^3p_{ij}}{d\theta^3} = 0,$$

we obtain for the expectations of nB_0, nB_1, nB_2, the results

$$\left.\begin{aligned} n\mathscr{E}(B_0) &= n\left\{\sum_i P_i\sum_j\frac{dp_{ij}}{d\theta}\right\}_{\theta_0} = 0, \\ n\mathscr{E}(B_1) &= -n\left\{\sum_i P_i\sum_j\frac{1}{p_{ij}}\left(\frac{dp_{ij}}{d\theta}\right)^2\right\}_{\theta_0} = -i(\theta_0)n, \\ n\mathscr{E}(B_2) &= n\left\{\sum_i P_i\sum_j\left[\frac{2}{p_{ij}^2}\left(\frac{dp_{ij}}{d\theta}\right)^3 - \frac{3}{p_{ij}}\frac{dp_{ij}}{d\theta}\frac{d^2p_{ij}}{d\theta^2}\right]\right\}_{\theta_1} = nK(\theta_1), \end{aligned}\right\} \tag{17}$$

where $i(\theta)$ is obviously finite and positive. We now prove that the variates $n_{ij}n^{-1}$ converge in probability to their expectations; from this it will follow that although the $n_{ij}n^{-1}$ are not independent, B_0, B_1, B_2, which are linear functions of these variates, also jointly converge in probability to their expectations.

In order to prove the convergence in probability of the variates $n_{ij}n^{-1}$ to their expectations P_ip_{ij}, we shall evaluate their variances using a method described in Fréchet (1952, p. 73). A transition frequency n_{ij} is treated as the sum of n variates $X_{ij}^{(r)}$, so that

$$n_{ij} = \sum_{r=1}^{n} X_{ij}^{(r)},$$

where $X_{ij}^{(r)}$ takes the values 1 or 0 at the rth transition, depending on whether this transition is or is not from the state E_i to the state E_j. It is clear that

$$\begin{aligned} \mathscr{E}(X_{ij}^{(r)}) &= P_ip_{ij}, \\ \mathscr{V}(X_{ij}^{(r)}) &= \mathscr{E}(X_{ij}^{(r)})^2\} - \{\mathscr{E}(X_{ij}^{(r)})\}^2 \\ &= P_ip_{ij} - P_i^2p_{ij}^2 \quad (r = 1, 2, \ldots, n), \\ \text{and} \qquad \mathscr{C}(X_{ij}^{(r)}X_{ij}^{(t)}) &= \mathscr{E}(X_{ij}^{(r)}X_{ij}^{(t)}) - \mathscr{E}(X_{ij}^{(r)})\mathscr{E}(X_{ij}^{(t)}) \\ &= P_ip_{ij}^2p_{jj}^{(t-r-1)} - P_i^2p_{ij}^2 \quad (r,t = 1, 2, \ldots, n;\ t>r), \end{aligned}$$

where $p_{ji}^{(t-r-1)}$ is the probability that a transition from the state E_j to the state E_i occur in $t-r-1$ steps, and is the (j,i)th element of the matrix $\mathbf{p}^{t-r-1}$, the $(t-r-1)$th power of the matrix $\mathbf{p}$.

The variance σ_{ij}^2 of n_{ij} is therefore given by

$$\begin{aligned}\sigma_{ij}^2 &= \mathscr{V}\left(\sum_{r=1}^{n} X_{ij}^{(r)}\right)\\ &= \sum_{r=1}^{n} \mathscr{V}(X_{ij}^{(r)}) + 2\sum_{t>r} \mathscr{C}(X_{ij}^{(r)} X_{ij}^{(0)})\\ &= n(P_i p_{ij} - P_i^2 p_{ij}^2) + 2P_i p_{ij}^2 \sum_{t>r} (p_{ji}^{(t-r-1)} - P_i),\end{aligned}$$

where there are $\frac{1}{2}n(n-1)$ terms of the form $(p_{ji}^{(t-r-1)} - P_i)$. Now, since the chain is ergodic, and therefore stationary, it is known that for all $t-r-1 \geqslant n_0$, where n_0 is some finite positive integer, the terms

$$|p_{ji}^{(t-r-1)} - P_i|$$

converge to zero at least as fast as the terms of a convergent geometric progression with a ratio q where $0<q<1$ depends on n_0. If, therefore, we write the variance σ_{ij}^2 as

$$\sigma_{ij}^2 = n(P_i p_{ij} - P_i^2 p_{ij}^2) + 2P_i p_{ij}^2 \sum_{k=0}^{n-2} (n-k-1)(p_{ji}^{(k)} - P_i),$$

we see that $\sigma_{ij}^2 n^{-1}$ will clearly converge to

$$\lim_{n\to\infty} \sigma_{ij}^2 n^{-1} = P_i p_{ij} - P_i^2 p_{ij}^2 + 2P_i p_{ij}^2 s_{ji} = A,$$

where A is some value independent of n, and $s_{ji} = \sum_{k=0}^{\infty} (p_{ji}^{(k)} - P_i)$.

We now prove that for ergodic chains, the limit A must be non-zero. To do this, we use a theorem given in Fréchet (1952, pp. 86–8) which applies to the frequencies n_i of (9), giving the number of times in a realization that the system is in state E_i. The variances $\sigma_i^2 = \mathscr{V}(n_i)$ of these can be shown, by a method similar to that used for the n_{ij} above, to be such that $\lim_{n\to\infty} \sigma_i^2 n^{-1}$ is some finite value independent of n. The theorem proves that for an ergodic chain of the type we consider, $\lim_{n\to\infty} \sigma_i^2 n^{-1}$ cannot be zero.

In order to apply this result to our transition frequencies n_{ij} we redefine our system of states in the following fashion, assuming first that the p_{ij} are all non-zero. We define a system of s^2 states E_{ij} $(i,j = 1,2,\ldots,s)$, in which the system will be in state E_{ij} when there is a transition from state E_i to state E_j of the original system; the new stochastic matrix for this system will be

$$\Upsilon = \begin{pmatrix} \begin{Bmatrix} p_{11} \cdots p_{1s} & 0 \ldots\ldots & \ldots\ldots & \ldots\ldots 0 \\ 0 \ \ldots\ 0 & p_{21} \cdots p_{2s} & 0 \ldots\ldots & \ldots\ldots 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0 \ldots\ldots & \ldots\ldots & \ldots\ldots 0 & p_{s1} \cdots p_{ss} \end{Bmatrix} \\ \begin{Bmatrix} \cdot \end{Bmatrix} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \begin{Bmatrix} \end{Bmatrix} \end{pmatrix},$$

If the original stochastic matrix is written as

$$\mathbf{p} = (\boldsymbol{\alpha}_1 \boldsymbol{\alpha}_2 \dots \boldsymbol{\alpha}_s) = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_s \end{pmatrix},$$

where $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_i$ are respectively column and row vectors, then $\boldsymbol{\gamma}$ can be written

$$\boldsymbol{\gamma} = \begin{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{\beta}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \boldsymbol{\beta}_s \end{pmatrix} \\ \begin{Bmatrix} & & \end{Bmatrix} \\ \dots\dots\dots\dots\dots\dots \\ \begin{Bmatrix} & & \end{Bmatrix} \end{pmatrix}.$$

It is clear that n_{ij} will now indicate the number of times in a realization of the chain that the system is in the state E_{ij}; in other words, the n_{ij} in the new system are the analogues of the n_i in the original system. All that remains to be proved is that if the original system is ergodic, so also will the new system; this is intuitively obvious, and can be easily shown by powering the matrix $\boldsymbol{\gamma}$ as follows. If we write for the nth power of the matrix $\mathbf{p}$

$$\mathbf{p}^n = (\boldsymbol{\alpha}_1^{(n)} \boldsymbol{\alpha}_2^{(n)} \dots \boldsymbol{\alpha}_s^{(n)}),$$

it is seen directly that on multiplying the matrix $\boldsymbol{\gamma}$ by itself, we obtain

$$\boldsymbol{\gamma}^2 = \left(p_{11} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_1 \end{pmatrix} \dots p_{1s} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_1 \end{pmatrix} p_{21} \begin{pmatrix} \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_2 \end{pmatrix} \dots p_{2s} \begin{pmatrix} \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_2 \end{pmatrix} \dots\dots p_{s1} \begin{pmatrix} \boldsymbol{\alpha}_s \\ \vdots \\ \boldsymbol{\alpha}_s \end{pmatrix} \dots p_{ss} \begin{pmatrix} \boldsymbol{\alpha}_s \\ \vdots \\ \boldsymbol{\alpha}_s \end{pmatrix} \right),$$

and similarly, since for $n \to \infty$, $\lim_{n\to\infty} \boldsymbol{\alpha}_1^{(n)} = \begin{pmatrix} P_1 \\ \vdots \\ P_1 \end{pmatrix}$,

then $$\boldsymbol{\gamma}^{n+1} = \left(p_{11} \begin{pmatrix} \boldsymbol{\alpha}_1^{(n)} \\ \vdots \\ \boldsymbol{\alpha}_1^{(n)} \end{pmatrix} \dots p_{ss} \begin{pmatrix} \boldsymbol{\alpha}_s^{(n)} \\ \vdots \\ \boldsymbol{\alpha}_s^{(n)} \end{pmatrix} \right) \to \begin{pmatrix} p_{11}P_1 & \dots & p_{ss}P_s \\ \vdots & & \vdots \\ p_{11}P_1 & \dots & p_{ss}P_s \end{pmatrix}.$$

It follows therefore by Fréchet's theorem that $\lim_{n\to\infty} \sigma_{ij}^2 n^{-1} \neq 0$. If a certain p_{ij} is zero, then in our new system the state E_{ij} must be eliminated, since no transition into or from this state is possible. It can be verified, however, that the results above will hold equally well in such a case. We now conclude from Tchebyshev's theorem that the variates $n_{ij}n^{-1}$ converge in probability to their expectations since

$$\Pr\{|\, n_{ij}n^{-1} - P_i p_{ij} \,| \geqslant \epsilon\} \leqslant \sigma_{ij}^2/n^2\epsilon^2 \leqslant A/n\epsilon^2.$$

We may express the joint convergence in probability of B_0, B_1, B_2 to their expectations in the form

$$\Pr\{|\, B_0 \,| < \delta^2,\ B_1 < -\tfrac{1}{2}i(\theta_0),\ |\, B_2 \,| < 2\,|\, K(\theta_1) \,|\} > 1 - \epsilon,$$

where δ, ϵ are any two small positive numbers. Let F denote the set of points for which this inequality holds; then for every point in F, we have that

$$B_0 + \tfrac{1}{2}B_2\delta^2 < (1 + |\, K(\theta_1) \,|)\,\delta^2$$

and

$$B_1\delta < -\tfrac{1}{2}i(\theta_0)\,\delta.$$

If $\theta = \theta_0 \pm \delta$ are two such points in F, then from (15) we have for these points

$$\frac{1}{n}\frac{dL}{d\theta} = B_0 \pm \delta B_1 + \tfrac{1}{2}\delta^2 B_2.$$

Provided that δ is so defined that

$$\delta < \tfrac{1}{2}i(\theta_0)\,(1 + |\,K(\theta_1)\,|)^{-1},$$

then for $\theta = \theta_0 + \delta$,

$$\frac{1}{n}\frac{dL}{d\theta} < (1 + |\,K(\theta_1)\,|)\{\delta - \tfrac{1}{2}i(\theta_0)\,(1 + |\,K(\theta_1)\,|)^{-1}\}\,\delta < 0,$$

whereas for $\theta = \theta_0 - \delta$,

$$\frac{1}{n}\frac{dL}{d\theta} > (1 + |\,K(\theta_1)\,|)\{\delta + \tfrac{1}{2}i(\theta_0)\,(1 + |\,K(\theta_1)\,|^{-1}\}\,\delta > 0.$$

We see from (6) that $dL/d\theta$ is a continuous function of θ in R, and it follows therefore that the likelihood equation (7) has a root in $\theta_0 \pm \delta$ with probability tending to 1. This establishes the consistency of the estimator T of θ.

(2) *Asymptotic normality of maximum-likelihood estimator*

Let the consistent solution of the likelihood equation be T; then from equation (15), we have that

$$\frac{1}{n}\left(\frac{dL}{d\theta}\right)_T = \frac{1}{n}\left(\frac{dL}{d\theta}\right)_{\theta_0} + (T-\theta_0)\frac{1}{n}\left(\frac{d^2L}{d\theta^2}\right)_{\theta_0} + \tfrac{1}{2}(T-\theta_0)^2\frac{1}{n}\left(\frac{d^3L}{d\theta^3}\right)_{\theta_1} = 0,$$

or

$$B_0 + (T-\theta_0)\,B_1 + \tfrac{1}{2}(T-\theta_0)^2 B_2 = 0,$$

so that

$$(T-\theta_0) = -B_0\{B_1 + \tfrac{1}{2}(T-\theta_0)\,B_2\}^{-1}.$$

We have seen that the expectation of B_0 is zero; its variance can be evaluated, again using the method described in Fréchet (1952, p. 73):

$$\mathscr{V}(B_0) = \mathscr{V}\left\{\sum_{i,j} n_{ij} n^{-1}\left(\frac{1}{p_{ij}}\frac{dp_{ij}}{d\theta}\right)_T\right\}$$

$$= \frac{1}{n^2}\mathscr{E}\left\{\sum_{i,j}\left(\frac{n_{ij}}{p_{ij}}\frac{dp_{ij}}{d\theta}\right)_T\right\}^2. \tag{18}$$

We treat each n_{ij} as the sum of the n variates $X_{ij}^{(r)}$ which take the values 1 or 0 at each of the n transitions $r = 1, \ldots, n$, depending on whether these transitions are or are not from the state E_i to the state E_j.

We then have

$$\mathscr{E}\left\{\sum_{i,j}\frac{n_{ij}}{p_{ij}}\frac{dp_{ij}}{d\theta}\right\}^2 = \mathscr{E}\left\{\sum_r\left(\sum_{i,j}\frac{X_{ij}^{(r)}}{p_{ij}}\frac{dp_{ij}}{d\theta}\right)\right\}^2$$

$$= \sum_r \mathscr{E}\left\{\sum_{i,j}\frac{X_{ij}^{(r)}}{p_{ij}}\frac{dp_{ij}}{d\theta}\right\}^2 + 2\sum_{r<t}\mathscr{E}\left\{\sum_{i,j}\frac{X_{ij}^{(r)}}{p_{ij}}\frac{dp_{ij}}{d\theta}\sum_{l,m}\frac{X_{lm}^{(t)}}{p_{lm}}\frac{dp_{lm}}{d\theta}\right\}$$

$$= \sum_r \mathscr{E}\left\{\sum_{i,j}\frac{(X_{ij}^{(r)})^2}{p_{ij}^2}\left(\frac{dp_{ij}}{d\theta}\right)^2 + \sum_{i,j}{\sum_{l,m}}'\frac{X_{ij}^{(r)}}{p_{ij}}\frac{X_{lm}^{(r)}}{p_{lm}}\frac{dp_{ij}}{d\theta}\frac{dp_{lm}}{d\theta}\right\}$$

$$+ 2\sum_{r<t}\left\{\sum_{i,j}\sum_{l,m}\mathscr{E}\left(\frac{X_{ij}^{(r)}}{p_{ij}}\frac{X_{lm}^{(t)}}{p_{lm}}\frac{dp_{ij}}{d\theta}\frac{dp_{lm}}{d\theta}\right)\right\},$$

where $\sum_r$, $\sum_{i,j}$, $\sum_{l,m}$ indicate summation over all possible values $r = 1, \ldots, n$; i, j and $l, m = 1, \ldots, s$, but where $\sum'_{i,j}$ $\sum'_{l,m}$ indicate summation over those values of i, j and $l, m = 1, \ldots, s$, for which at least one of the conditions $i \neq l, j \neq m$ holds. Since for these values of i, j and l, m, we have that $X_{ij}^{(r)} X_{lm}^{(r)} = 0$, and further $\sum_m \frac{dp_{lm}}{d\theta} = 0$ for all l, then this is equal to

$$\sum_{i,j} \mathscr{E}\left\{\sum_r \frac{X_{ij}^{(r)}}{p_{ij}^2}\left(\frac{dp_{ij}}{d\theta}\right)^2\right\} + 2\sum_{r<t}\left\{\sum_{i,j}\sum_{l,m}(P_i p_{ij} p_{jl}^{(t-r-1)} p_{lm})\left(\frac{1}{p_{ij}p_{lm}}\frac{dp_{ij}}{d\theta}\frac{dp_{lm}}{d\theta}\right)\right\} = n\sum_{i,j}\frac{P_i}{p_{ij}}\left(\frac{dp_{ij}}{d\theta}\right)^2.$$

It follows that $$\mathscr{V}(B_0) = -n^{-1}\mathscr{E}(B_1) = n^{-1} i(\theta_0),$$

so that

$$(T-\theta_0)\sqrt{\{i(\theta_0)\, n\}} = \frac{\dfrac{B_0}{\sqrt{\{n^{-1}i(\theta_0)\}}}}{\left\{\dfrac{B_1}{-i(\theta_0)} + \dfrac{(T-\theta_0)\,B_2}{-2i(\theta_0)}\right\}}$$

has zero expectation and unit variance, for the denominator converges to 1 in probability, and the numerator has zero expectation and unit variance. If we can further show that E_0 is asymptotically normally distributed, then T will be asymptotically normally distributed with expectation θ_0 and variance $\{i(\theta_0)\, n\}^{-1}$.

Now from (16) we see that B_0 is a linear function of the n_{ij}, so that it is sufficient to prove that as $n \to \infty$, these are jointly normally distributed. This has been proved, for the ergodic chains considered, by Bartlett (1951), who in his work also implied some of the properties of maximum-likelihood estimators developed in detail in this paper. The proof is briefly outlined below; one difference, however, is that it is no longer necessary to postulate $\ln \mu_1(\mathbf{t}) \neq 0$, since it was shown in §III that for the chains considered, this must be so for some $\mathbf{t}$.

Following Montroll (1947) and Bartlett (1951), we define the moment generating function of the n_{ij} as

$$M(\mathbf{t}) = \mathscr{E}(\exp \sum_{i,j} t_{ij} n_{ij})$$
$$= \mathscr{E}(\exp \sum_{i,j} t_{ij} \sum_r X_{ij}^{(r)}).$$

If $\mathbf{R}$ is the square matrix whose transpose $\mathbf{R}'$ is defined by (10), this is

$$M(\mathbf{t}) = \mathbf{1}'\mathbf{R}^n\mathbf{P},$$

where $\mathbf{1}'$ denotes the row vector $(1, \ldots, 1)$, and $\mathbf{P}$ is the column vector of the stationary probabilities P_i. If $\mathbf{R}$ has the s latent roots $\mu_1(\mathbf{t}), \ldots, \mu_s(\mathbf{t})$, we have seen that for small $\mathbf{t}$. $\mu_1(\mathbf{t})$ is the single dominant root such that $\mu_1(\mathbf{0}) = 1$. Following Frazer, Duncan & Collar (1947, §4·15), we have that Sylvester's theorem gives for $n \to \infty$,

$$\mathbf{R}^n \sim \{\mu_1(\mathbf{t})\}^n\, \mathbf{Z}_0(\mu_1),$$

where the matrix $\mathbf{Z}_0(\mu_1)$ is finite. It follows that

$$M(\mathbf{t}) \sim \{\mu_1(\mathbf{t})\}^n\, \mathbf{1}'\mathbf{Z}_0(\mu_1)\,\mathbf{P},$$

so that $$\ln M(\mathbf{t}) \sim n \ln \mu_1(\mathbf{t}),$$

a non-zero value, since for an ergodic chain, for some small $\mathbf{t}$ it has been shown that $\mu_1(\mathbf{t}) \neq 1$.

$M(\mathbf{t})$ is thus asymptotically equivalent to the moment generating function of a sum of n identical independent vector variates. Since for an ergodic chain, the n_{ij} have variances of order n so that $\sigma_{ij}^2 n^{-1}$ are finite and non-zero, they will tend to simultaneous normality as $n \to \infty$ by the Central Limit Theorem for $n = \sum_{i,j} n_{ij}$ variates. It follows that B_0 will be asymptotically normally distributed.

Finally, a general convergence theorem of Cramér (1946) enables us to conclude that $(T-\theta_0)\sqrt{\{i(\theta_0)n\}}$ is asymptotically normally distributed with zero expectation and unit variance; this is that if a variate ξ_n with distribution function $f_n(x)$ tending to $f(x)$ and a variate η_n converging in probability to a positive constant c as $n \to \infty$ exist, then the distribution function of $\xi_n \eta_n^{-1}$ tends to $f(cx)$.

Since
$$\{i(\theta_0)n\}^{-1} = \left\{\mathscr{E}\left(\left|\frac{d^2L}{d\theta^2}\right|_{\theta_0}\right)\right\}^{-1}$$
is the minimum possible variance, we see clearly that the consistent solution of the maximum-likelihood equation is fully efficient.

IV. Sufficiency conditions

A further general theorem of maximum-likelihood estimation, proved under certain conditions by Cramér (1946, § 33·2) for continuous distributions, applies equally with minor modifications to the case of the Markov chain. It is that:

If a sufficient estimator T of the parameter θ exists, any solution of the likelihood equation will be a function of T. In obtaining the maximum-likelihood estimator T of θ, the parameter defining the transition probabilities $p_{ij}(\theta)$ of a Markov chain, it may therefore be of interest to determine the form of the transition probabilities which admit a sufficient statistic for θ.

Koopman (1936) and Pitman (1936) have given the general form of continuous distributions admitting a sufficient statistic, but although it has been obvious that a similar form exists for discrete distributions, no account of the proof for this appears to be available. We briefly outline the proof, use it to obtain the form of the probabilities $p_1(\theta), \ldots, p_k(\theta)$, in the particular case of the multinomial distribution, and finally generalize our results in the case of the simple ergodic Markov chain.

(1) *The form of the likelihood function admitting a sufficient estimator T of θ. for discrete parent distributions with constant variate intervals*

We consider those discrete distributions $p(x, \theta)$ for which the variate x is defined at equal intervals. We can then, without loss of generality, assume the interval to be unity so that x takes consecutive integral values in any specific range, finite or infinite; this includes the standard discrete distributions for which x is a frequency, as well as others for which x may take positive or negative integral values.

Let $x_1, \ldots, x_n$ be n independent observations of a variate x with the discrete distribution $p(x, \theta)$, where x takes a set of consecutive integral values in a given range. The probability of this sample is
$$\phi(x_1, \ldots, x_n; \theta) = p(x_1, \theta) \ldots p(x_n, \theta).$$
If this is to admit a sufficient estimator T of θ, then the factorizability condition
$$\phi(x_1, \ldots, x_n; \theta) = f(x_1, \ldots, x_n) F(T, \theta)$$

must hold, or for the likelihood function $L = \ln \phi$,

$$L = \ln f(x_1, \ldots, x_n) + \ln F(T, \theta). \tag{19}$$

Assuming that the functions $p(x_i, \theta)$, $F(T, \theta)$ are differentiable with respect to θ, we obtain the equation

$$\sum_i \frac{\partial}{\partial \theta} \{\ln p(x_i, \theta)\} = \frac{\partial}{\partial \theta} \{\ln F(T, \theta)\} = G(T, \theta). \tag{20}$$

Since this holds for all values of θ within the range allowed, a particular value θ_0 lying in this range will give, when substituted in this equation,

$$u = \sum_i u(x_i) = g(T), \tag{21}$$

which connects T with the statistic

$$u = \sum_i u(x_i) = \sum_i \left(\frac{\partial}{\partial \theta} \{\ln p(x_i, \theta)\} \right)_{\theta = \theta_0}.$$

Now since sufficiency is a property which holds for all values of the sample $x_1, \ldots, x_n$, we may allow x_i, a particular discrete sample value, to increase by 1, while the remaining $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, remain unchanged. We can then difference equation (21) with respect to x_i to obtain

$$\Delta_{x_i} u = \Delta_{x_i} u(x_i) = \Delta_{x_i} g(T).$$

Similarly for equation (20)

$$\Delta_{x_i} \frac{\partial}{\partial \theta} \ln p(x_i, \theta) = \Delta_{x_i} G(T, \theta),$$

so that for all values of i, we have the equation

$$\frac{\Delta_{x_i} \dfrac{\partial}{\partial \theta} \ln p(x_i, \theta)}{\Delta_{x_i} u(x_i)} = \frac{\Delta_{x_i} G(T, \theta)}{\Delta_{x_i} g(T)}. \tag{22}$$

Now if $G(T, \theta)$ and $g(T)$ are assumed differentiable with respect to T, we have that

$$\begin{aligned} \Delta_{x_i} G(T, \theta) &= G(T(x_1, \ldots, x_{i-1}, x_i + 1, x_{i+1}, \ldots, x_n), \theta) - G(T(x_1, \ldots, x_i, \ldots, x_n), \theta) \\ &= G(T_1, \theta) - G(T_0, \theta) \\ &= (T_1 - T_0) \left| \frac{\partial G}{\partial T} \right|_{T_2} \\ &= (\Delta_{x_i} T) \left| \frac{\partial G}{\partial T} \right|_{T_2}, \end{aligned}$$

where T_2 is some function $T_2(x_1, \ldots, x_i, \ldots, x_n; \theta)$ of the sample values and of θ such that $T_0 \leqslant T_2 \leqslant T_1$. Further, since the estimator T is a symmetrical function of the x_i, the function T_2 will be the same in the case of every differencing with respect to each x_i, $i = 1, \ldots, n$. Similarly,

$$\Delta_{x_i} g(T) = (\Delta_{x_i} T) \left| \frac{\partial g}{\partial T} \right|_{T_2(x_1, \ldots, x_i, \ldots, x_n; \theta_0)},$$

where

$$T_0 \leqslant T_2(x_1, \ldots, x_i, \ldots, x_n; \theta_0) \leqslant T_1.$$

It follows that equation (22) can be written for all values of i as

$$\frac{\Delta_{x_i} \dfrac{\partial}{\partial \theta} \ln p(x_i, \theta)}{\Delta_{x_i} u(x_i)} = \frac{\left| \dfrac{\partial G}{\partial T} \right|_{T_2(x_1, \ldots, x_n; \theta)}}{\left| \dfrac{dg}{dT} \right|_{T_2(x_1, \ldots, x_n; \theta_0)}},$$

which is equal to a fixed value, not depending on i. Hence

$$\frac{\Delta_{x_1}\frac{\partial}{\partial\theta}\ln p(x_1,\theta)}{\Delta_{x_1}u(x_1)}=\frac{\Delta_{x_2}\frac{\partial}{\partial\theta}\ln p(x_2,\theta)}{\Delta_{x_2}u(x_2)}=\ldots=\frac{\Delta_{x_n}\frac{\partial}{\partial\theta}\ln p(x_n,\theta)}{\Delta_{x_n}u(x_n)},$$

and this can only be so if these are all equal to a function of θ only, $K_1(\theta)$. It follows from (22) that

$$\Delta_{x_i}G(T,\theta)=K_1(\theta)\,\Delta_{x_i}g(T),$$

so that

$$G(T,\theta)=K_1(\theta)\,g(T)+K_2(\theta).$$

On integrating this with respect to θ, we obtain

$$\ln F(T,\theta)=\Lambda_1(\theta)\,g(T)+\Lambda_2(\theta)+h(T),$$

so that the likelihood (19) is of the form

$$L=\ln f(x_1,\ldots,x_n)+\Lambda_1(\theta)\,g(T)+\Lambda_2(\theta), \tag{23}$$

and the probability of the sample is of the form

$$\phi(x_1,\ldots,x_n;\,\theta)=f(x_1,\ldots,x_n)\exp\left[\Lambda_1(\theta)\,g(T)+\Lambda_2(\theta)\right].$$

(2) *Form of the probabilities $p_i(\theta)$ for a multinomial distribution admitting a sufficient estimator T of θ*

Consider the k mutually exclusive events $E_1, E_2, \ldots, E_k$, with non-zero probabilities $p_1(\theta), \ldots, p_k(\theta)$ such that $\sum_{i=1}^{k} p_i(\theta)=1$. In a total of n independent trials in which there are x_1 occurrences of the events E_1, x_2 of the event E_2, ..., and x_k of the event E_k, the probability of the sample of frequencies $x_1, x_2, \ldots, x_k$, is

$$\phi(x_1,\ldots,x_k;\,\theta)=p_1^{x_1}(\theta)\ldots p_k^{x_k}(\theta).$$

For this to admit a sufficient estimator, the likelihood $L=\ln\phi$ must be of the form (23), so that

$$L=\sum_{i=1}^{k}x_i\ln p_i(\theta)=\ln f(x_1,\ldots,x_k)+\Lambda_1(\theta)\,g(T)+\Lambda_2(\theta).$$

We note in this case that $\Lambda_2(\theta)$ involves the number of trials $n=\sum_{i=1}^{k}x_i$ so that it can be written

$$\Lambda_2(\theta)=\sum_{i=1}^{k}x_i\lambda_2(\theta),$$

where $\lambda_2(\theta)$ is a function of θ only. Also from the form of the likelihood, we see that the first term can only be of the form

$$\ln f(x_1,\ldots,x_k)=\sum_{i=1}^{k}A_i x_i.$$

We can therefore write the likelihood function admitting a sufficient estimator T of θ as

$$L=\sum_i x_i\ln p_i(\theta)=\sum_i A_i x_i+\Lambda_1(\theta)\,g(T)+\lambda_2(\theta)\sum_i x_i, \tag{24}$$

which holds for all values of n, and of the frequency values $x_1, \ldots, x_n$. If, in a total of $n+1$ independent trials, the frequency of each event $E_1, E_2, \ldots, E_{i-1}, E_{i+1}, \ldots, E_k$ apart from the

event E_i remains unchanged, while the event E_i occurs x_i+1 times, then we may difference equation (24) with respect to x_i, to obtain

$$\ln p_i(\theta) = A_i + \Lambda_1(\theta)\,\Delta_{x_i} g(T) + \lambda_2(\theta),$$

or

$$\{\Lambda_1(\theta)\}^{-1}\{\ln p_i(\theta) - A_i - \lambda_2(\theta)\} = \Delta_{x_i} g(T).$$

That is, a function of θ only equals a function of the sample values x_i, which is possible only if both equal a constant K_i.

It follows that

$$\begin{aligned} p_i(\theta) &= \exp[K_i \Lambda_1(\theta) + \lambda_2(\theta) + A_i] \\ &= \alpha_i \exp[K_i \Lambda_1(\theta) + \lambda_2(\theta)] \quad (i = 1, \ldots, k), \end{aligned} \tag{25}$$

where $A_i = \ln \alpha_i$, is the form of the probability allowing a sufficient estimator T of θ.

A condition to which the probabilities $p_i(\theta)$ are subject is

$$\sum_{i=1}^{k} p_i(\theta) = 1,$$

so that in our case, from (25)

$$\sum_i \alpha_i \exp[K_i \Lambda_1(\theta) + \lambda_2(\theta)] = 1,$$

or

$$\sum_i \alpha_i \exp[K_i \Lambda_1(\theta)] = \exp[-\lambda_2(\theta)].$$

An alternative way of writing the probabilities $p_i(\theta)$ is therefore

$$\begin{aligned} p_i(\theta) &= \alpha_i \exp[K_i \Lambda_1(\theta)] \{\textstyle\sum_i \alpha_i \exp[K_i \Lambda_1(\theta)]\}^{-1} \\ &= \alpha_i \lambda_1^{K_i}(\theta) \{\textstyle\sum_i \alpha_i \lambda_1^{K_i}(\theta)\}^{-1}, \end{aligned} \tag{26}$$

where

$$\lambda_1(\theta) = \exp[\Lambda_1(\theta)].$$

We can now tell immediately in any practical case whether it is possible to obtain a sufficient estimator or not by an examination of the form of the multinomial probabilities. Three simple illustrations from trinomial distributions are given as examples.

Example IV 2·1. The trinomial distribution with probabilities

$$p_1 = \theta\{\theta + \theta^2 + 2\theta^3\}^{-1}, \quad p_2 = \theta^2\{\theta + \theta^2 + 2\theta^3\}^{-1}, \quad p_3 = 2\theta^3\{\theta + \theta^2 + 2\theta^3\}^{-1},$$

of the form (26) with α_i equal to 1, 1, 2 respectively and K_i equal to 1, 2, 3, where $\lambda_1(\theta) = \theta$, will give an estimator of θ which is sufficient.

Example IV. 2·2. Probabilities of the type

$$p_1 = \theta, \quad p_2 = 2\theta, \quad p_3 = 1 - 3\theta,$$

which can be written in the form (25)

$$p_i = \alpha_i \exp[K_i \ln \theta(1 - 3\theta)^{-1} + \ln(1 - 3\theta)],$$

where α_i equal 1, 2, 1, and K_i equal 1, 1, 0, respectively, will also give an estimator of θ which is sufficient.

Example IV. 2·3. However, probabilities of the form

$$p_1 = \tfrac{1}{4}(2 + \theta), \quad p_2 = \tfrac{1}{2}(1 - \theta), \quad p_3 = \tfrac{1}{4}\theta,$$

which occur in genetics, cannot be written in the forms (25) or (26), and can be verified to admit no sufficient estimator of θ.

(3) *Form of the transition probabilities $p_{ij}(\theta)$ for a simple ergodic Markov chain admitting a sufficient estimator T of θ*

Consider the realization of the simple ergodic Markov chain with s possible states $E_1, \ldots, E_s$ which results in the observed sequence S of the $n+1$ states $E_i, E_j, \ldots, E_k, E_l$, in this order. Assuming the initial state E_i to be fixed,* the probability distribution of the sequence S is given by (2), and the likelihood can be written as equation (3). An ergodic chain (see footnote, § III) is such that for any row i of the transition matrix, and for all values of j,

$$0 \leqslant p_{ij}(\theta) < 1,$$

so that, in particular, for $i = j$ no state E_i is an absorption state; some, though not all, $p_{ij}(\theta)$ in a row may equal zero, subject to the usual conditions (1) that

$$\sum_{j=1}^{s} p_{ij}(\theta) = 1 \quad (i = 1, 2, \ldots, s).$$

Now if the chain admits a sufficient estimator T of θ, we see that it may be written, much in the same way as for the multinomial case,

$$L = \sum_{i,j} n_{ij} \ln p_{ij}(\theta) = \sum_{i,j} A_{ij} n_{ij} + \Lambda_1(\theta)\, g(T) + \lambda_2(\theta) \sum_{i,j} n_{ij}, \tag{27}$$

where the A_{ij} are constants, $\Lambda_2(\theta)$ is equal to the function $\lambda_2(\theta) \sum_{i,j} n_{ij}$ in this case, and $\sum_{i,j} n_{ij} = n$.

Before proceeding, as in the similar case of the multinomial, to increase the number of trials in the realization of the chain by one, so as to allow this equation to be differenced with respect to the n_{ij} in order to obtain the non-zero values of the $p_{ij}(\theta)$, we must consider some difficulties which this method presents. Since to any zero transition probability $p_{ij}(\theta)$ there will necessarily correspond a zero value of the transition frequency n_{ij}, neither will appear in the likelihood function (27). We are concerned with the n_{ij} corresponding to non-zero values of $p_{ij}(\theta)$, and differencing occurs only with respect to these; however, owing to the set of relations (8) for the sums of transition frequencies, it is not always possible in increasing the number of trials in a realization of the chain by 1, from $n+1$ to $n+2$, to increase each of the n_{ij} associated with a non-zero transition probability $p_{ij}(\theta)$ singly by 1, while leaving the values of the remaining n_{ij} unchanged.

As a simple example to illustrate this point, we consider the case of a chain with two states E_1 and E_2, for which a possible realization of three trials, starting with the fixed state E_1, results in the sequence E_1, E_1, E_2. Assuming all the transition probabilities to

* We are specifically considering the case of sufficiency for a distribution conditional on the given initial state E_i. In the general case where any state E_i may occur initially, the likelihood is

$$L = \sum_{i=1}^{s} x_i \ln a_i + \sum_{i,j} n_{ij} \ln p_{ij},$$

where the variates x_i are 0 or 1 according to whether the state E_i is or is not the initial state. In addition to the usual conditions for the n_{ij}, the x_i are subject to the condition that

$$\sum_{i=1}^{s} x_i = 1.$$

This more difficult case will not be considered.

have non-zero values, we see immediately that the transition frequency matrix with elements n_{ij} is

$$\mathbf{n} = \begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix};$$

a realization of four trials starting with E_1 will give, among others, the following transition frequency matrices:

$$\begin{pmatrix} 2 & 1 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

for the sequences E_1, E_1, E_1, E_2; E_1, E_1, E_2, E_1; and E_1, E_1, E_2, E_2. Here we see that n_{11}, n_{21}, n_{22} have singly been increased by 1, while the remaining n_{ij} remain constant; it is, however, in no way possible, starting with the given matrix $\mathbf{n}$, to increase n_{12} by 1, while leaving n_{11}, n_{21}, n_{22} unchanged.

In order to resolve this difficulty, we use the fact that for an ergodic chain, providing the number of trials $n+1$ is sufficiently large in a realization starting with the given state E_i, it is possible to end with any one of the states $E_1, \ldots, E_s$. In our particular case, assuming a sufficiently large number of trials, the realization of the chain leads to the sequence S which starts with the fixed state E_i and ends with the state E_l; a further trial may result in the sequence S' with an additional state E_r, for all r for which $p_{lr}(\theta) \neq 0$. The probability of this sequence is

$$\phi(S') = p_{ij}(\theta) \ldots p_{kl}(\theta)\, p_{lr}(\theta),$$

and the likelihood function $L' = \ln \phi(S')$ is identical with $L = \ln \phi(S)$ except that the transition frequency n_{lr} in it is increased to $n_{lr}+1$. This allows us to difference the equation (27) with respect to n_{lr} for all r for which $p_{lr} \neq 0$, in order to obtain

$$\ln p_{lr}(\theta) = A_{lr} + \Lambda_1(\theta)\, \Delta_{n_{lr}} g(T) + \lambda_2(\theta),$$

where $\Delta_{n_{lr}} g(T)$ is some function of the transition frequencies n_{ij}. Now since

$$\{\Lambda_1(\theta)\}^{-1} \{\ln p_{lr} - \lambda_2(\theta) - A_{lr}\} = \Delta_{n_{lr}} g(T),$$

that is, a function of θ equals a function of the n_{ij}, then both must equal a constant K_{lr}. This allows us to write for all r for which $p_{lr} \neq 0$,

$$p_{lr}(\theta) = \alpha_{lr} \exp\left[K_{lr} \Lambda_1(\theta) + \lambda_2(\theta)\right],$$

where $A_{lr} = \ln \alpha_{lr}$. If we further admit zero values of α_{lr}, we may accept this equation as the general form for both zero and non-zero values of p_{lr} $(r = 1, 2, \ldots, s)$. Since, for a sufficiently large number of trials in a realization of the chain starting with the given state E_i, the final state E_l may be such that l can take any of the values $1, 2, \ldots, s$, it follows that we may write as the general form of the transition probabilities $p_{ij}(\theta)$ admitting a sufficient estimator T of θ, the equation

$$p_{ij}(\theta) = \alpha_{ij} \exp\left[K_{ij} \Lambda_1(\theta) + \lambda_2(\theta)\right]. \tag{28}$$

This general form of the $p_{ij}(\theta)$ has been obtained for sufficiently large values of $n = \sum_{i,j} n_{ij}$; this means that in this case the necessary condition that a sufficient estimator of θ exists is that the $p_{ij}(\theta)$ be of form (28). It is easily verified that the sufficient condition also follows; that if the $p_{ij}(\theta)$ are of the form (28), the estimator of θ is sufficient. This, however, holds generally, irrespective of the size of n. It is of interest to note that if we restrict ourselves to the case where all the $p_{ij}(\theta)$ are non-zero, all values of $n \geqslant 1$ are sufficiently large for the

necessary condition to hold, since any final state E_l may be reached in a realization of one or more trials starting with the given state E_i. This means that for the ergodic chain with non-zero $p_{ij}(\theta)$, a sufficient estimator of θ exists for any number of trials $n \geqslant 1$, if and only if the $p_{ij}(\theta)$ are of the form (28).

The transition probabilities (28) are subject to the condition (1) that $\sum_{j=1}^{s} p_{ij} = 1$, so that

$$\sum_{j=1}^{s} \alpha_{ij} \exp[K_{ij}\Lambda_1(\theta)] = \exp[-\lambda_2(\theta)],$$

or

$$\sum_{j=1}^{s} \alpha_{ij} \exp[K_{ij}\Lambda_1(\theta)] = \sum_{j=1}^{s} \alpha_{rj} \exp[K_{rj}\Lambda_1(\theta)] \quad (i, r = 1, 2, \ldots, s). \tag{29}$$

If we write $\lambda_1(\theta) = \exp[\Lambda_1(\theta)]$, we see that the $p_{ij}(\theta)$ can also be put in the form

$$p_{ij}(\theta) = \alpha_{ij}\lambda_1^{K_{ij}}(\theta)\{\sum_j \alpha_{ij}\lambda_1^{K_{ij}}(\theta)\}^{-1}, \tag{30}$$

where $\{\sum_j \alpha_{ij}\lambda_1^{K_{ij}}(\theta)\}^{-1}$ has the same value for all rows $i = 1, 2, \ldots, s$.

Two possible cases now exist for those exponents among $K_{i1}, K_{i2}, \ldots, K_{is}$, associated with the non-zero probabilities among $p_{i1}, p_{i2}, \ldots, p_{is}$, in the ith row of the transition probability matrix $\mathbf{p}$; the first in which all such exponents are distinct, and the second in which only some such exponents are distinct. For simplicity, we assume that in the ith row, all transition probabilities $p_{ij}(\theta)$ $(j = 1, 2, \ldots, s)$ are non-zero; the slightly more general case where some $p_{ij}(\theta)$ may be zero follows without difficulty.

Consider the first case, where for the particular row i there are s exponents K_{ij}, all distinct, associated with the s non-zero transition probabilities $p_{ij}(\theta)$ $(j = 1, 2, \ldots, s)$; then, from equation (29), it follows that for any other row r,

$$\sum_j \alpha_{ij}\lambda_1^{K_{ij}} = \sum_j \alpha_{rj}\lambda_1^{K_{rj}}. \tag{31}$$

This is possible only if the exponents K_{ij} and K_{rj} $(j = 1, 2, \ldots, s)$ have the same s distinct values, though these may be arranged in different orders. This means that there is only a single set of s distinct values

$$\alpha_{i1}\lambda_1^{K_{i1}}\{\sum_j \alpha_{ij}\lambda_1^{K_{ij}}\}^{-1}, \quad \ldots, \quad \alpha_{is}\lambda_1^{K_{is}}\{\sum_j \alpha_{ij}\lambda_1^{K_{ij}}\}^{-1},$$

for the transition probabilities of the matrix $\mathbf{p}$, and these must appear, possibly in different orders, in every row of the matrix.

Two simple examples for Markov chains with two and three states respectively will illustrate the previous points:

Example IV. 3·1. The transition probability matrix of the form

$$\mathbf{p} = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}$$

will always provide a sufficient estimator of θ, for we may write $p_{ij}(\theta)$ $(j = 1, 2)$ in the form (28) as

$$p_{1j}(\theta) = \alpha_{1j} \exp[K_{1j} \ln \theta(1-\theta)^{-1} + \ln(1-\theta)],$$

where the values of α_{11}, α_{12} and K_{11}, K_{12} are 1, 1 and 1, 0 respectively. The second row consists of the same transition probabilities in a different order.

Example IV. 3·2.

$$\mathbf{p} = \begin{pmatrix} 2\theta(2\theta+\theta^2+\theta^3)^{-1} & \theta^2(2\theta+\theta^2+\theta^3)^{-1} & \theta^3(2\theta+\theta^2+\theta^3)^{-1} \\ \theta^3(2\theta+\theta^2+\theta^3)^{-1} & 2\theta(2\theta+\theta^2+\theta^3)^{-1} & \theta^2(2\theta+\theta^2+\theta^3)^{-1} \\ \theta^2(2\theta+\theta^2+\theta^3)^{-1} & \theta^3(2\theta+\theta^2+\theta^3)^{-1} & 2\theta(2\theta+\theta^2+\theta^3)^{-1} \end{pmatrix}$$

also provides a sufficient estimator of θ, for in (30) we have $\lambda_1(\theta) = \theta$, and the values of α_{1j} and K_{1j} are 2, 1, 1 and 1, 2, 3 respectively. The second and third rows consist of the same values for the transition probabilities, but in different orders.

Consider now the second case for the exponents. Here, for the particular row i, the s exponents K_{ij} are not all distinct; let the first k exponents ($k<s$) be identical

$$K_{i1} = K_{i2} = \ldots = K_{ik} = K$$

and the remaining $K_{i,k+1}, \ldots, K_{is}$ be distinct, so that there are altogether $s-k+1$ distinct values for the exponents, and the transition probabilities for row i are therefore

$$\left.\begin{aligned} p_{ij}(\theta) &= \alpha_{ij}\lambda_1^K\Big\{\sum_{j=1}^{s}\alpha_{ij}\lambda_1^{K_{ij}}\Big\}^{-1} \quad (j=1,2,\ldots,k), \\ p_{ij}(\theta) &= \alpha_{ij}\lambda_1^{K_{ij}}\Big\{\sum_{j=1}^{s}\alpha_{ij}\lambda_1^{K_{ij}}\Big\}^{-1} \quad (j=k+1,\ldots,s). \end{aligned}\right\} \tag{32}$$

(The slightly more general case of several groups of exponents with identical values, and the remaining exponents distinct

$$\begin{aligned} K_{i1} &= K_{i2} = \ldots = K_{ik} = K^{(1)}, \\ K_{i,k+1} &= K_{i,k+2} = \ldots = K_{im} = K^{(2)}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ K_{iu} &\neq K_{i,u+1} \neq \ldots \neq K_{is}, \end{aligned}$$

presents no difficulty different from those met in the simpler case mentioned, and will not be considered.) Now since for any row r equation (31) holds, there are $s-k+1$ distinct values of K_{rj} which are identical with the distinct values of the K_{ij}, so that in every row of the stochastic matrix $\mathbf{p}$ we have, apart from coefficients, transition probabilities of only the following $s-k+1$ distinct forms

$$\lambda_1^K\{\textstyle\sum_j \alpha_{ij}\lambda_1^{K_{ij}}\}^{-1}, \quad \lambda_1^{K_{i,k+1}}\{\textstyle\sum_j \alpha_{ij}\lambda_1^{K_{ij}}\}^{-1}, \quad \ldots, \quad \lambda_1^{K_{is}}\{\textstyle\sum_j \alpha_{ij}\lambda_1^{K_{ij}}\}^{-1}, \tag{33}$$

appearing in various arrangements. The s coefficients associated with these $s-k+1$ distinct forms for the transition probabilities may also differ from row to row, but are subject to the condition which follows from (31) that for all rows $r = 1, 2, \ldots, s$, the sums of the coefficients for each distinct form in (33) must equal the fixed values

$$\sum_{j=1}^{k}\alpha_{ij}, \quad \alpha_{i,k+1}, \quad \ldots, \quad \alpha_{is}, \tag{34}$$

respectively.

Two simple examples of this case for Markov chains with three states follow.

Example IV. 3·3.

$$\mathbf{p} = \begin{pmatrix} 2\theta(3\theta+\theta^3)^{-1} & \theta(3\theta+\theta^3)^{-1} & \theta^3(3\theta+\theta^3)^{-1} \\ \frac{1}{3}\theta^3(3\theta+\theta^3)^{-1} & \frac{2}{3}\theta^3(3\theta+\theta^3)^{-1} & 3\theta(3\theta+\theta^3)^{-1} \\ \frac{1}{4}\theta(3\theta+\theta^3)^{-1} & \theta^3(3\theta+\theta^3)^{-1} & \frac{1}{4}\theta(3\theta+\theta^3)^{-1} \end{pmatrix},$$

where in equation (32), $\lambda_1(\theta) = 0$, and the two distinct forms for the transition probabilities appearing in each row are

$$\theta(3\theta+\theta^3)^{-1}, \quad \theta^3(3\theta+\theta^3)^{-1}.$$

The values of the coefficients for the form $\theta(3\theta+\theta^3)^{-1}$ vary from row to row, subject to the condition (34) that their sum is always equal to 3, while those for the form $\theta^3(3\theta+\theta^3)^{-1}$ have a sum always equal to 1.

Example IV. 3·4.

$$\mathbf{p} = \begin{pmatrix} \frac{1}{2}\theta & 1-\theta & \frac{1}{2}\theta \\ \frac{1}{4}\theta & \frac{3}{4}\theta & 1-\theta \\ \frac{1}{2}(1-\theta) & \frac{1}{2}(1-\theta) & \theta \end{pmatrix},$$

where in equation (32), $\lambda_1(\theta) = \theta(1-\theta)^{-1}$, and the two distinct forms for the transition probabilities appearing in each row are

$$\{\theta(1-\theta)^{-1}\}^{K_{ij}}(1-\theta), \quad K_{ij} = 0, 1.$$

Again the values of the coefficients for θ vary from row to row, subject to condition (34) that their sum is always equal to 1, while those for $(1-\theta)$ also have a sum equal to 1.

I am greatly indebted to Prof. P. A. P. Moran for suggesting the field of parameter estimation in simple Markov chains as a subject for research, and for his useful critical comments throughout all stages of the work. I should also like to thank the referee for several helpful suggestions.

REFERENCES

BARTLETT, M. S. (1951). The frequency goodness of fit test for probability chains. *Proc. Camb. Phil. Soc.* **47**, 86–95.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

FRAZER, R. A., DUNCAN, W. J. & COLLAR, A. R. (1947). *Elementary Matrices*. New York: Macmillan.

FRÉCHET, M. (1952). *Traité du Calcul des Probabilités et de ses Applications*, tome **2**, fasc. III. Gauthier-Villars.

HUZURBAZAR, V. S. (1948). The likelihood equation, consistency and maxima of the likelihood function. *Ann. Eugen., Lond.*, **14**, 185–200.

KOOPMAN, B. O. (1936). On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.* **39**, 399–409.

MONTROLL, E. W. (1947). On the theory of Markoff chains. *Ann. Math. Statist.* **18**, 18–36.

PITMAN, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. *Proc. Camb. Phil. Soc.* **32**, 567–79

RAO, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. New York: John Wiley.

SIGNIFICANCE TESTS FOR DISCRIMINANT FUNCTIONS AND LINEAR FUNCTIONAL RELATIONSHIPS

By E. J. WILLIAMS

Division of Mathematical Statistics, C.S.I.R.O., Melbourne

I. Introduction

In previous papers (Bartlett, 1951; Williams, 1952*a*) certain exact tests for the adequacy of a hypothetical discriminant function were derived. Later papers (Williams, 1952*b*, 1952*c*, 1953) showed how these tests could be applied in a number of situations of practical usefulness. The first object of the present paper is to extend the work in the above-mentioned papers and to show how the results obtained may be interpreted in terms of multiple linear regression. The calculations may indeed be carried out in the manner of a covariance analysis. The second object is to develop, along the same lines, exact tests for an assumed linear relationship among variables; this is a problem which has been discussed in various contexts by Koopmans (1937), Tintner (1945, 1946, 1950), Geary (1948, 1949), Bartlett (1948), Anderson (1951) and others. Since the question of determining underlying relationships has been given considerable attention in the literature from a number of different points of view, the opportunity is taken also to discuss and to attempt to unify the different approaches made—the use of information provided by instrumental variates, by grouping of the data, and by higher moments.

The reason for discussing discriminant functions and functional relationships in the same paper is because the two problems are really different aspects of the same problem. This has been well demonstrated by Geary (1948). If a single discriminant function is assumed adequate to describe differences among a number of p-variate populations, this assumption is equivalent to assuming that there exist $p-1$ linear relations among the means for the p variates; the means then lie on a line. In general, postulating that the differences among the populations are described by r discriminant functions is equivalent to postulating $p-r$ linear relationships (provided always that the number of populations considered is not less than p). The quantity r, the number of dimensions in which the population means lie, may be called the rank of the populations, and $p-r$ the degeneracy.

Thus the test for a single linear relationship is equivalent to the test for the adequacy of $p-1$ discriminant functions. In deriving significance tests for either a discriminant function or a linear relationship, the same principle is applied, though the function being tested has a different role in the two cases, and thus enters differently into the tests. In the simple bivariate case, the test for a linear relationship is exactly the same as the test for the discriminant function which is orthogonal to it.

The problems of this paper have been framed above in terms of an analysis of variance model, for testing the significance of differences between populations. This has been done in order to link them with those discussed in the earlier work (Williams, 1952*a*, *b*). A more general specification is in terms of a regression model, wherein the interrelationships between a set of p variates and another set of q variates are investigated. In such a model a discriminant function is better described as a canonical variate. Throughout the remainder of

this paper the problems will be posed in terms of the regression model but interpreted also in terms of analysis of variance. Though both specifications are formally equivalent it is easier sometimes to work with one and sometimes with the other.

Bartlett (1951) has shown that the different aspects of the adequacy of a discriminant function may be tested by means of a factorization of a certain determinantal ratio. A general calculus of such factorizations is developed, and it is shown how it can be applied in a number of the tests described in this paper.

The method of treatment used earlier (Williams, 1952*a*) in deriving exact significance tests was to frame the questions asked of the data somewhat differently from those usually posed. A hypothetical discriminant function being proposed, the question is whether this function is concordant with the data. Since the hypothetical discriminant function leads to a sufficient statistic for the unknown population parameter, tests can be derived which are independent of the value of the parameter. These tests are the counterparts of those based on the latent roots (largest or smallest) of a matrix, but have the advantage of being exact, even in small samples. The same approach is adopted in this paper, in deriving tests both for discriminant functions and functional relationships.

The most important practical results given in the present paper are that the tests previously given for the adequacy of a hypothetical discriminant function, and the tests now given for a functional relationship, can be reduced to tests derived by a covariance analysis of the data. The computations are thereby simplified because it is not necessary to determine the latent roots of a matrix equation, the test functions being expressed either as ordinary adjusted sums of squares in a covariance analysis, or as the ratios of determinants of sums of squares and products. If, however, the 'best' discriminant function yielded by the sample (or the 'best' linear relationship) is to be evaluated, it is still necessary to evaluate the largest (smallest) latent root and the corresponding latent vector.

II. NOTATION, TERMINOLOGY AND PRELIMINARY RESULTS

Since this paper covers a rather wide field it has been thought desirable to set out clearly the notation to be used throughout. While an attempt has been made to use notation in conformity with previous work on the subject, some departures have been made in the interests of clarity.

We consider two sets of variates,

$$X_i \quad (i = 1, 2, \ldots, p),$$
$$Y_j \quad (j = 1, 2, \ldots, q),$$

measured on each of $n+1$ individuals. The X_i will be assumed to have a non-singular joint normal distribution, with linear regression on the Y_j. Formally, this is equivalent to $q+1$ populations with the same p-variate normal distribution of values about the means in each, so that the variation in the sample may be analysed into the q degrees of freedom between groups and the $n-q$ within groups.

The discrimination problem is to determine from a set of data, either that linear function or set of linear functions of the X_i which have greatest correlation with the Y_j, or the adequacy of a set of given discriminant functions. On the other hand, the problem of determining functional relationships is to find the set of linear functions, if any, of the X_i which are uncorrelated with the Y_j, or to test a given set of such functions for their correlation with

the Y_j. Linear functions which are uncorrelated with the Y_j, either by hypothesis or in the sample, will be called null functions, and their sample values null variates.

Although from the formal correlation point of view there is a symmetry between the X_i and the Y_j, the tests of significance developed here are always of a function (either a discriminant function or a null function) of the variables of one of the sets (in this paper the X_i), and so are not symmetrical in the two sets. In this respect the present approach differs from that based on the canonical correlations alone, in which the magnitudes of the canonical correlations are used to decide the number of discriminant functions (or, in other words, the rank and degeneracy of the population correlations). Reiersøl (1945) emphasizes the distinction between the two sets of variates by designating the Y_j 'instrumental' variates to distinguish them from the 'investigational' variates X_i.

The canonical variates of the two sets will be denoted by x_i, y_i, and the corresponding squared canonical correlations by θ_i. The hypothetical discriminant function will be denoted by ξ, and its squared multiple correlation with the Y_j, or discriminant ratio, by λ.

The tests of significance for a discriminant function have been shown in earlier papers to depend only on the original discriminant ratios θ_i and the new discriminant ratios after the effect of ξ has been eliminated by covariance, which will be denoted by ϕ_i.

It will be shown similarly in this paper that the tests for a linear functional relationship, that is, for a given null function, depend on sets of original and new latent roots; the new latent roots which will be denoted by θ'_i are, however, defined differently from those used in discriminant analysis.

The following notation will be used:

Sums of squares and products:

t_{hi}	sum of products of X_h and X_i
p_{ij}	sum of products of X_i and Y_j
u_{jk}	sum of products of Y_j and Y_k
b_{hi}	sum of products of X_h and X_i for regression or between-group line of analysis of variance
w_{hi}	sum of products of X_h and X_i for residuals from regresssion or within groups.
Also $t_{\xi i}$	sum of products of ξ and X_i
$t_{\xi\xi}$	sum of squares of ξ

and similarly for other sums of squares and products with the suffix ξ.

The corresponding matrices will be denoted by capitals. Then the following results are readily obtained, relating the different expressions occurring in the tests of significance:

$$B+W=T,$$
$$B=PU^{-1}P', \qquad |B|/|T| = \Pi\theta_i,$$
$$W=T-PU^{-1}P', \quad |W|/|T| = \Pi(1-\theta_i).$$

The matrix of sums of squares and products of all the $p+q$ variates will be denoted by S:

$$S=\begin{bmatrix} T & P \\ P' & U \end{bmatrix}.$$

Armed with these results, we are in a position to set out expressions in terms either of the regression model or the analysis of variance model.

In the derivation of tests of significance by means of covariance analysis, use is made of sums of squares and products of the X_i after adjustment by certain covariance variates. Since there seems to be no consistent terminology for these quantities, we propose to use the term 'adjusted' only for quantities which are adjusted without any loss of degrees of freedom, and the term 'reduced' for quantities for which the degrees of freedom are reduced by the number of covariance variates. Thus, the analysis of variance of any X_i would give the following partition of degrees of freedom:

	Degrees of freedom
Regression	q
Residual	$n-q$
Total	n

After adjustment by r covariance variates, the analysis would give

	Degrees of freedom
Regression, adjusted	q
Residual, reduced	$n-q-r$
Total, reduced	$n-r$

The adjusted regression sum of squares may, if required, be partitioned as follows:

	Degrees of freedom
Difference of regressions on covariance variates	r
Regression, reduced	$q-r$
Regression, adjusted	q

III. THE FACTORIZATION OF DETERMINANTAL RATIOS

The overall likelihood criterion for testing the rank of the population from which a sample has been drawn is, in terms of the sample latent roots, equal to

$$\prod_{i=1}^{p}(1-\theta_i).$$

In terms of determinants of sums of squares and products it may also be expressed as

$$|W|/|T| = |T - PU^{-1}P'|/|T| = \frac{|S|}{|T||U|}, \tag{1}$$

and so may be represented as a determinantal ratio without calculation of the latent roots. The criterion may be expressed as the ratio of two determinants of order p, with $n-q$ and n degrees of freedom respectively; and since the X_i and the Y_j enter symmetrically into the expression, it may also be expressed as the ratio of two determinants of order q, with degrees of freedom $n-p$ and n. It will accordingly be denoted

$$(n; p, q)$$

to indicate its dependence on the three sets of degrees of freedom. In particular, for $p = 1$, the ratio of a residual sum of squares with $n-q$ degrees of freedom to a total sum of squares with n degrees of freedom is denoted $(n; 1, q)$.

Now if s is an integer less than q, a sum of squares with q degrees of freedom can be partitioned into two independent sums of squares with s and $q-s$ degrees of freedom. Corresponding to this partition, the ratio of sums of squares with $n-q$ and n degrees of freedom can be factorized into two independent factors: one with $n-s$ and n degrees of freedom, the other with $n-q$ and $n-s$ degrees of freedom; so that

$$(n; 1, q) = (n; 1, s)(n-s; 1, q-s). \tag{2}$$

In exactly the same way, determinantal ratios may be factorized, giving, for example,

$$(n; p, q) = (n; p, s)(n-s; p, q-s), \tag{3}$$

or, more generally,

$$= (n; p, s)(n-s; r, q-s)(n-r-s; p-r, q-s). \tag{4}$$

These results are the basis of the factorizations of likelihood criteria given by Bartlett (1951), and will be applied repeatedly throughout this paper.

As in univariate analysis, the first factor in (3) is seen to be the likelihood criterion for the simple effect of the s variables, the remaining $q-s$ being ignored, while the second factor corresponds to the partial effect of the $q-s$ variables after the elimination of the first s. The partial factor is accordingly the one which is to be used in significance tests. Just as in univariate analysis, there are always two alternative factorizations, depending on which set out of the q variables is to be eliminated and which is to be the subject of test. Thus

$$\begin{aligned}(n; p, q) &= (n; p, s)(n-s; p, q-s) \\ &= (n-q+s; p, s)(n; p, q-s).\end{aligned} \tag{5}$$

When $p = 1$, the tests for the partial factors may be simultaneously set out in the form of an analysis of variance:

Effect	Degrees of freedom	Sum of squares
Partial s variates	s	$1-(n; 1, s)$
Partial $q-s$ variates	$q-s$	$1-(n; 1, q-s)$
Residual	$n-q$	$(n; 1, q)$
Total	n	1

The possibility of thus arranging the tests arises from the fact that (for example)

$$\frac{1-(n;\,1,s)}{(n;\,1,q)} = \frac{1-(n-q+s;\,1,s)}{(n-q+s;\,1,s)}$$

$$= \frac{s}{n-q} F_{(s,\,n-q)}. \tag{6}$$

The moments of the likelihood criterion may be determined by means of this factorization. Taking $s = 1$ and factorizing successively, we have

$$(n;\,p,q) = (n;\,p,1)\,(n-1;\,p,1)\,(n-2;\,p,1)\ldots(n-q+1;\,p,1).$$

It is readily shown that the tth moment about zero of $(n;\,p,1)$ is

$$E[(n;\,p,1)^t] = \frac{\Gamma[\frac{1}{2}(n-p)+t]\,\Gamma(\frac{1}{2}n)}{\Gamma[\frac{1}{2}(n-p)]\,\Gamma(\frac{1}{2}n+t)},$$

so that the moments of the likelihood criterion are given by

$$E[(n;\,p,q)^t] = \prod_{i=0}^{q-1} \frac{\Gamma[\frac{1}{2}(n-p-i)+t]\,\Gamma[\frac{1}{2}(n-i)]}{\Gamma[\frac{1}{2}(n-p-i)]\,\Gamma[\frac{1}{2}(n-i)+t]} \tag{7}$$

(cf. Bartlett (1938), equation (26)).

An important particular case is the limiting one when n tends to infinity. Writing

$$(p,q) = \lim_{n\to\infty} (n;\,p,q)^n,$$

we find that the factorization (3) becomes

$$(p,q) = (p,s)\,(p,q-s), \tag{8}$$

and it can be shown that $-\log(p,q)$ is distributed as a sum of squares with pq degrees of freedom (cf. Williams, 1952*a*, p. 26).

When n tends to infinity, the two factorizations given above are identical, and are equivalent to a simple partition of a sum of squares—as is, indeed, evident from the analysis set out above for the particular case $p = 1$.

An application. As a simple example of the application of these methods we consider the case where r of the population latent roots are known to be unity, and the vanishing of the remaining population roots is under test.

If r population roots are unity, it follows that $\theta_1, \theta_2, \ldots, \theta_r$ are all unity, and that the corresponding canonical variates $x_1, x_2, \ldots, x_r$ are equal (apart from possible changes of sign) to $y_1, y_2, \ldots, y_r$, the canonical variates of the second set. Therefore, in order to examine the residual variation of the x_i, the p variates of the first set may be replaced by any $p-r$ of them not linearly dependent on the first r of them. Each of these variates has $n-q$ residual degrees of freedom. For their regression on the y_i, the variates are restricted by the condition of orthogonality with the first r x_i and hence with the first r y_i, so that the regression has but $q-r$ degrees of freedom. Accordingly, the residual likelihood criterion for testing the existence of the non-unit latent roots, namely,

$$\prod_{i=r+1}^{p} (1-\theta_i) \tag{9}$$

is a $(n-r; p-r, q-r)$ variable. In particular, if $r=p-1$ $(p \leqslant q)$, the single sample latent root θ_p may be tested by an F-test:

$$1-\theta_p = (n-p+1;\ 1, q-p+1),$$

$$F = \frac{(n-q)\theta_p}{(q-p+1)(1-\theta_p)}, \tag{10}$$

with $q-p+1$ and $n-q$ degrees of freedom.

Corresponding results hold when $q<p$.

These results have practical application where, even if the first r population canonical correlations are not unity, the sample latent roots are close to unity. Then the elimination of the corresponding sample canonical variates leads to a residual likelihood criterion (9) with approximately the distribution given. Such a result has already been discussed by Bartlett (1947) and others.

IV. Review and extension of previous work

Bartlett (1951) derived tests of significance for the goodness of fit of a single hypothetical discriminant function, in terms of its direction, and of departures of the data from collinearity. These results generalized some results of Williams (1952*a*) for the particular case $p = 2$ (two variates) and $q = 2$ (two degrees of freedom between groups). However, two points were not clarified in this work. First, while Williams used for test functions the partial factors, Bartlett suggested that the use of partial factors was not necessary in these tests. It should be clear from the discussion in § III that the partial factors are appropriate for significance tests.

Secondly, it was stated by Bartlett that the general test functions could be expressed in terms of the original latent roots and the new latent roots after adjustment for the hypothetical discriminant function, but the expressions were not given. These results, which are given below, show the relationship of the exact tests to the approximate tests based on the individual ordered latent roots. Apart from this, they are of more analytic than practical interest, since the calculation of all the latent roots, before and after adjustment, is laborious. Moreover, in § V of this paper, simplified computational procedures based on the analysis of covariance will be given, which make possible the ready application of the exact tests without the calculation of latent roots.

Expression of test functions in terms of canonical correlations

For this section, we express all results in terms of the canonical variates x_i, y_i and the canonical correlations. Let the hypothetical discriminant function which is under test be

$$\xi = \Sigma m_i x_i, \quad \text{with} \quad \Sigma m_i^2 = 1,$$

so that the corresponding discriminant ratio is

$$\lambda = \Sigma m_i^2 \theta_i.$$

The first factorization of the overall likelihood criterion, representing the elimination of the effect of the hypothetical discriminant function, is

$$\prod_i (1-\theta_i) = (1-\lambda) \prod_j (1-\phi_j),$$

corresponding to $(n; p, q) = (n; 1, q)(n-1; p-1, q)$, and leading by the elimination of the factor $1-\lambda$ for the hypothetical discriminant function to the residual likelihood criterion $\prod_j (1-\phi_j)$.

Bartlett (1951) gives two factorizations of the residual likelihood criterion, providing tests for simple direction and partial collinearity (equation (7) of his paper) and for partial direction and simple collinearity (equation (20) of his paper). In our notation, these factorizations may be written

$$\prod_{\substack{j \\ (n-1;\,p-1,\,q)}} (1-\phi_j) = \prod (1-\theta_i)/(1-\lambda),$$

$$= \underbrace{\left[\frac{1-\Sigma m_i^2\theta_i^2/\lambda}{1-\lambda}\right]}_{\substack{\text{Simple direction} \\ (n-1;\,p-1,\,1)}} \underbrace{\left[\frac{\Pi(1-\theta_i)}{1-\Sigma m_i^2\theta_i^2/\lambda}\right]}_{\substack{\text{Partial collinearity} \\ (n-2;\,p-1,\,q-1)}} \tag{11}$$

$$= \underbrace{\left[\frac{\lambda}{(1-\lambda)\Sigma \dfrac{m_i^2\theta_i}{1-\theta_i}}\right]}_{\substack{\text{Partial direction} \\ (n-q;\,p-1,\,1)}} \underbrace{\left[\frac{\Pi(1-\theta_i)}{\lambda}\Sigma\frac{m_i^2\theta_i}{1-\theta_i}\right]}_{\substack{\text{Simple collinearity} \\ (n-1;\,p-1,\,q-1)}}. \tag{12}$$

The degrees of freedom for these factors agree with the values given by Bartlett. Since the first factor of (12) has unity for one of its degrees of freedom, it may be tested by an ordinary F-test.

It now remains to express the m_i^2 in terms of the original and new latent roots. From Williams (1952a, p. 20) we have, in the present notation, the equation for the ϕ_j

$$\sum_i \frac{m_i^2(1-\theta_i)}{\theta_i-\phi_j} = 0 \quad (j = 1, 2, \ldots, p-1). \tag{13}$$

On expressing this as an equation of degree $p-1$ in ϕ_j, we have

$$\phi_j^{p-1}\sum_i m_i^2(1-\theta_i) - \phi_j^{p-2}\sum_i m_i^2(1-\theta_i)(\sum_h \theta_h - \theta_i) + \ldots = 0,$$

i.e.
$$\phi_j^{p-1}(1-\lambda) - \phi_j^{p-2}[\sum_i \theta_i(1-\lambda) - \lambda + \sum_i m_i^2\theta_i^2] + \ldots = 0.$$

Hence
$$\sum_j \phi_j = \sum_i \theta_i - (\lambda - \sum_i m_i^2\theta_i^2)/(1-\lambda),$$

so that
$$(\lambda - \sum_i m_i^2\theta_i^2)/(1-\lambda) = \sum_i \theta_i - \sum_j \phi_j,$$

and the factors (11) become
$$\left[\frac{\sum_i \theta_i - \sum_j \phi_j}{\lambda}\right]\left[\frac{\lambda\prod_j (1-\phi_j)}{\sum_i \theta_i - \sum_j \phi_j}\right].$$

The second factor provides the test for collinearity. It takes its maximum value $\prod_2^p (1-\theta_i)$, when $\lambda = \theta_1$. The use of this product therefore overestimates the significance of departures from collinearity. When $p = 2$ it reduces to

$$\frac{\dfrac{(1-\theta_1)(1-\theta_2)}{1-\lambda}}{\dfrac{(1-\theta_1)(1-\theta_2)}{1-\lambda} + \dfrac{\theta_1\theta_2}{\lambda}} = [1 + v_2/(1-v_1)(1-v_2)]^{-1} \quad \text{(notation of Williams, 1952}a\text{)},$$

and is thus a function of the variance ratio previously given for testing collinearity.

When $q = 2$ the second factor of (11) becomes

$$\frac{(1-\theta_1)(1-\theta_2)}{\theta_1+\theta_2-\phi_1-\phi_2} = 1-Q/P \quad \text{(notation of Williams, 1952}a\text{)}$$
$$= [1+v_2/(1-v_1)(1-v_2)]^{-1},$$

which again corresponds to the test function given in Williams (1952*a*, p. 26).

For the alternative factorization it is convenient to write

$$\Theta = \theta/(1-\theta),$$
$$\Phi = \phi/(1-\phi),$$
$$\Lambda = \lambda/(1-\lambda).$$

Then equation (13) becomes

$$\sum_i \frac{m_i^2}{\Theta_i - \Phi_j} = 0 \quad (j = 1, 2, \ldots, p-1),$$

so that
$$\Phi_j^{p-1} - \Phi_j^{p-2} \sum_i m_i^2 (\sum_h \Theta_h - \Theta_i) + \ldots = 0.$$

Hence
$$\sum m_i^2 \Theta_i = \sum_i \Theta_i - \sum_j \Phi_j,$$

and the factors (12) become

$$\left[\frac{\Lambda}{\sum_i \Theta_i - \sum_j \Phi_j}\right]\left[\frac{\prod_j (1-\phi_j)}{\Lambda}(\sum_i \Theta_i - \sum_j \Phi_j)\right]. \qquad (12')$$

The first factor provides the test for direction; being a $(n-q;\ p-1,\ 1)$ variable it may be tested by means of an F-test with $p-1$ and $n-p-q+1$ degrees of freedom:

$$F = \frac{n-p-q+1}{p-1}\frac{(1-(n-q;\ p-1,\ 1))}{(n-q;\ p-1,\ 1)}.$$

When λ is close to θ_1, so that the hypothetical discriminant function differs little in direction from the first canonical variate, the first factor of (12) is approximately λ/θ_1. This provides a justification of the approximate F-test

$$F = \frac{n-p-q+1}{p-1}\frac{(\theta_1-\lambda)}{\lambda}.$$

When $p = 2$ the first factor becomes

$$\frac{(1-\theta_1)(1-\theta_2)}{(1-\lambda)(1-\theta_1\theta_2/\lambda)} = 1-v_1,$$

the test function previously given.

When $q = 2$ the first factor reduces to

$$(1-v_1)(1-v_2),$$

again equivalent to the test function given previously.

The two factorizations may be shown in closer relationship by means of the following representations, which make use of the fact that $(1-\phi_j) = (1+\Phi_j)^{-1}$:

$$\prod_j (1-\phi_j) = \left[\frac{\sum_i \theta_i - \sum_j \phi_j}{\lambda}\right]\left[\frac{\lambda \prod_j (1-\phi_j)}{\sum_i \theta_i - \sum_j \phi_j}\right],$$

$$\prod_j (1+\Phi_j) = \left[\frac{\sum_i \Theta_i - \sum_j \Phi_j}{\Lambda}\right]\left[\frac{\Lambda \prod_j (1+\Phi_j)}{\sum_i \Theta_i - \sum_j \Phi_j}\right].$$

V. Derivation of tests for direction and collinearity by covariance analysis

It has been shown by Bartlett (1939) that a general test for the adequacy of a discriminant function is provided by the residual likelihood criterion after the elimination by covariance of the effect of the function. A point that has been overlooked, however, is that by a simple extension of this method the separate tests for direction and collinearity of the function may be derived. Bartlett there deals with the simple case of discrimination between two groups ($q = 1$). In the treatment below we shall consider the regression model, in the general case with q independent (or instrumental) variates.

(a) *The case* $p = 2$

To introduce the method, we consider first two variates X_1 and X_2, and their correlation with the q variates $Y_1, Y_2, \ldots, Y_q$. Let the hypothetical discriminant function be

$$\xi = l_1 X_1 + l_2 X_2.$$

To test the adequacy of ξ, we consider the residual correlation of the X_i with the Y_j after the elimination of ξ by covariance. We may adjust either X_1, X_2 or any linear function of them, for its dependence on ξ; the same analysis is achieved in any case, since ξ is itself a linear function of X_1 and X_2.

For the overall test, the sum of squares, with q degrees of freedom, for regression on the Y_j after adjustment for ξ, is tested against the reduced residual sum of squares, with $n-q-1$ degrees of freedom. Now the adjusted regression sum of squares may be separated into two parts: (i) one with $q-1$ degrees of freedom, namely, the sum of squares for regression, reduced by the elimination of the effect of ξ, and (ii) one with 1 degree of freedom, representing the regression of the regression values on ξ.

In the analysis of variance model with q degrees of freedom between groups, the two parts of the adjusted sum of squares between groups are (i) the reduced sum of squares between groups, and (ii) the sum of squares for the difference of between-group and within-group regressions.

Then it turns out that the reduced sum of squares (i) gives the sum of squares for testing the independence of the adjusted X_i and Y_j (or the collinearity of populations in the analysis of variance model), while the sum of squares (ii) provides the test for direction of the proposed discriminant function. The tests are equivalent to those given in the formal analysis of variance presented previously (Williams, 1952*a*, p. 24).

This equivalence may be established as follows.

With the notation given in § II, we can express the determinantal ratios in terms of symmetric functions of the latent roots (discriminant ratios). Thus

$$|B|/|T| = \theta_1\theta_2,$$

$$|W|/|T| = (1-\theta_1)(1-\theta_2),$$

while, by definition,

$$b_{\xi\xi}/t_{\xi\xi} = \lambda.$$

Then the sums of squares for X_1, after adjustment by covariance for ξ, may be expressed in terms of determinantal ratios or latent roots as follows:

	Degrees of freedom	Sums of squares in terms of	
		Determinantal ratios	Latent roots
Difference of regressions (direction)	1	$l_2^2(\lvert T\rvert/t_{\xi\xi} - \lvert B\rvert/b_{\xi\xi} - \lvert W\rvert/w_{\xi\xi})$	$\frac{(\theta_1-\lambda)(\lambda-\theta_2)}{\lambda(1-\lambda)}\frac{l_2^2\lvert T\rvert}{t_{\xi\xi}}$
Between-groups reduced (collinearity)	$q-1$	$l_2^2\lvert B\rvert/b_{\xi\xi}$	$\frac{\theta_1\theta_2}{\lambda}\frac{l_2^2\lvert T\rvert}{t_{\xi\xi}}$
Within-groups reduced	$n-q-1$	$l_2^2\lvert W\rvert/w_{\xi\xi}$	$\frac{(1-\theta_1)(1-\theta_2)}{1-\lambda}\frac{l_2^2\lvert T\rvert}{t_{\xi\xi}}$
Total reduced	$n-1$	$l_2^2\lvert T\rvert/t_{\xi\xi}$	$\frac{l_2^2\lvert T\rvert}{t_{\xi\xi}}$

The right-hand column gives terms which are proportional to the quantities used in testing direction and collinearity against a residual term. This analysis further justifies the procedure given in the earlier papers, of testing the partial factors (i.e. terms corresponding to the terms given in the above covariance analysis) rather than the simple factors.

This approach to the tests for the adequacy of a hypothetical discriminant function greatly simplifies the computations. By this method it is not necessary to calculate the θ_i, nor even to determine the canonical variates x_i. A straightforward covariance analysis is all that is required. However, if what is required is not a test of the adequacy of a given discriminant function but the calculation of the most satisfactory discriminant function from the data, then the canonical analysis is still necessary.

(*b*) *The case* $p > 2$

The procedure outlined above may be generalized for greater values of p. We take the hypothetical discriminant function to be

$$\xi = l_1X_1 + l_2X_2 + \ldots + l_pX_p,$$

and consider the analysis of covariance of the X_i on ξ. Since ξ is a linear function of the X_i, the number of linearly independent variates after the elimination of ξ is reduced to $p-1$. As for the case $p = 2$, the sums of squares and products of the residual variates may be

separated into terms for direction and collinearity, the term for direction having a single degree of freedom. From the $p-1$ residual variates, a linear function may be made up in the same way as a discriminant function for a comparison for which the sum of squares for direction is relatively maximized. This sum of squares clearly has $p-1$ degrees of freedom. When the sum of squares is maximized relative to the reduced total sum of squares, we get the sum of squares for the simple direction effect (collinearity being ignored) (11). When it is maximized relative to the reduced sum of squares for residuals from regression on the Y_j, we get the sum of squares for the partial direction effect (12). The analysis will be demonstrated for the simple criterion for direction; the partial criterion may be similarly derived.

For any two of the variates X_h and X_i, the sum of products for direction may be derived by difference in the manner given above (when $p = 2$), but may be expressed more simply for the present purpose as

$$d_h d_i b_{\xi\xi} w_{\xi\xi}/t_{\xi\xi},$$

where

$$d_i = \left(\frac{b_{\xi i}}{b_{\xi\xi}} - \frac{w_{\xi i}}{w_{\xi\xi}}\right).$$

Now consider some linear compound of the X_i, say

$$\zeta = l'_1 X_1 + l'_2 X_2 + \ldots + l'_p X_p.$$

For this compound, the sum of squares for direction is

$$\left(\sum_{i=1}^{p} l'_i d_i\right)^2 b_{\xi\xi} w_{\xi\xi}/t_{\xi\xi},$$

while the total sum of squares is

$$\sum_1^p \sum_1^p l'_h l'_i t_{hi}.$$

To maximize the ratio of direction to total, which we denote by D, with respect to the l'_i, we have

$$\sum_1^p l'_h t_{hi} D t_{\xi\xi} = d_i \sum_1^p l'_h d_h b_{\xi\xi} w_{\xi\xi}.$$

The scale of the l'_i being arbitrary, we may so choose it that

$$D t_{\xi\xi} = \sum_1^p l'_h d_h b_{\xi\xi} w_{\xi\xi}, \tag{14}$$

whence

$$\sum_1^p l'_h t_{hi} = d_i. \tag{15}$$

If we write the typical element of the inverse of T as t^{hi}, then the solutions of (15) are given by

$$l'_i = \sum_1^p t^{hi} d_h.$$

Hence, from (14),

$$D = \sum_1^p \sum_1^p t^{hi} d_h d_i b_{\xi\xi} w_{\xi\xi}/t_{\xi\xi}.$$

It is to be noted that ζ is orthogonal to ξ. This may be obvious from its derivation; alternatively, the total sum of products of ζ and ξ is

$$\sum_1^p \sum_1^p l_h l'_i t_{hi} = \sum_1^p l_h d_h$$
$$= 0,$$

while the sum of products for direction is proportional to

$$\sum_1^p \sum_1^p l_h l'_i d_h d_i = 0.$$

The quantity D. being an invariant, may readily be expressed in terms of latent roots in order to show the equivalence of the present results with those given in § IV. Transforming to the canonical variates x_i, and putting as before

$$\xi = \Sigma m_i x_i,$$

we have

$$d_i = \frac{m_i(\theta_i - \lambda)}{\lambda(1-\lambda)},$$

$$D = \frac{\Sigma m_i^2(\theta_i - \lambda)^2}{\lambda(1-\lambda)}$$

$$= \frac{\Sigma m_i^2 \theta_i^2 - \lambda^2}{\lambda(1-\lambda)},$$

so that

$$1 - D = \frac{1 - \Sigma m_i^2 \theta_i^2/\lambda}{1-\lambda},$$

which agrees with (11).

To derive the test function for partial collinearity in terms of determinantal ratios, we make use of the factorization (11).

The residual likelihood criterion

$$\Pi(1-\theta_i)/(1-\lambda) = \frac{|W|}{|T|} \frac{t_{\xi\xi}}{w_{\xi\xi}}.$$

Hence the partial collinearity factor is

$$\frac{|W|}{|T|} \frac{t_{\xi\xi}}{w_{\xi\xi}} \Big/ (1-D) = \frac{|W|}{|T|} \frac{t_{\xi\xi}^2}{w_{\xi\xi}} \Big/ \left(t_{\xi\xi} - b_{\xi\xi} w_{\xi\xi} \sum_1^p \sum_1^p t^{hi} d_h d_i \right), \tag{16}$$

which is a $(n-2: p-1, q-1)$ variate.

In the same way, the partial direction factor may be found by maximizing the ratio of the sum of squares for direction to the reduced residual sum of squares. The maximized ratio is

$$\frac{t_{\xi\xi}}{t_{\xi\xi} + b_{\xi\xi} w_{\xi\xi} \Sigma w^{hi} d_h d_i}. \tag{17}$$

This is a $(n-q; p-1, 1)$ variate, so may be tested by the F-test:

$$F = \frac{n-p-q+1}{p-1} \frac{b_{\xi\xi} w_{\xi\xi}}{t_{\xi\xi}} \Sigma w^{hi} d_h d_i,$$

with $p-1$ and $n-p-q+1$ degrees of freedom.

It may be verified that these results agree with the general results given by Bartlett (1951) and the particular results given by Williams (1952*a*) for $p = 2, q = 2$.

VI. General remarks on the linear functional relationship

As has often been remarked, the linear functional relationship between two variates subject to error is different from either of the regression relationships between the two; the relationships are identical only when the independent variate in the regression relationship is

errorless. Also the two relationships have different applications. The regression relationship is the more generally useful, relating as it does to observed values; an important application is the prediction of either observed or 'true' values of one variate from observed values of the other. The functional relationship connects the 'true' values of the two variates; its use is in theoretical studies of underlying 'laws'. In most practical applications the regression and not the functional relationship is required. Since we are here considering only linear relationships, the word 'linear' will usually be omitted.

The regression relationships are based on the variation in both the 'true' values and the random errors to which they are subject, the functional relationship on the variation in the 'true' values alone. A little reflexion will show (and examination of the literature will confirm; see, for example, Haavelmo (1943)) that the functional relationship is therefore relevant only to a study of how the 'true' values of both variates are affected by some extraneous variate or variates; that is to say, the relationship shows what elements of the system are invariant under changes in conditions. It is not of interest to know the underlying relationship (if any) between two variates when each is affected only by random error; usually what is then wanted is one or other of the regression relationships.

It is well known that, when both the 'true' values and the errors are normally distributed, the functional relationship cannot be determined from data. Lindley (1947), Reiersøl (1950) and others have proved a number of theorems which show that the non-normality of one of the distributions is necessary for the estimation of the relationship to be possible. Thus it may be taken that functional relationships are not determinable from the internal analysis of a set of variates with normal distributions. Put in another way, since the first- and second-order sample cumulants summarize all the information in samples from normal populations, it follows that information beyond the first- and second-order cumulants of the distributions must be available if the functional relationship is to be determinable. This information may be present in the sample, if the distributions are not normal, for then the first- and second-order sample cumulants are no longer sufficient statistics; alternatively, it may be introduced through knowledge of the relationship of each variate with some extraneous variates. As mentioned above, it is only under such circumstances, when some such additional information exists, that the functional relationship is of interest; in other words, whenever functional relationships are of practical interest they are also determinable from data.

In passing it may be mentioned that Berkson's (1950) method of 'controlled' variables, as elucidated by Lindley (1953), is really a means of dealing with one variate as though it were errorless, and bringing the estimation of the functional relationship back to the estimation of a regression equation.

The foregoing discussion could be generalized for the determination of relationships among three or more variates, for which the same general considerations apply. We shall be concerned henceforth with the determination of functional relationships among a set of variates through their relationships with the variates of a second set. Reiersøl (1941, 1945) has termed the second set *instrumental* variates, to distinguish them from the first set of *investigational* variates. This approach has been used also by Geary (1943, 1949) and others.

Another method (Wald, 1940; Bartlett, 1949) for determining the functional relationship uses groupings of the values of the variates. Neyman & Scott (1951) have shown that only 'in very exceptional circumstances' does the method provide consistent estimates. Roughly

speaking, the method leads to consistent results provided the separation of values of one of the variates is sufficiently wide in the neighbourhood of the group limits, so that the grouping based on observed values is equivalent to a grouping based on 'true' values. Clearly, under these conditions, the differences between groups may be attributed to some extraneous variates, so that the method of grouping is then roughly equivalent to the method of instrumental variates.

Both methods have this in common, that they introduce extraneous variation affecting all the variables between which the relationship is found. The existence or assumption of this variation is basic to the methods.

In the next sections of this paper we give some exact tests for the existence of and for the constants in linear functional relationships, based on the use of instrumental variates. We shall assume for convenience that the investigational variates are subject to normally distributed errors, but make no particular assumptions about the distribution of the instrumental variates. Indeed, the instrumental variates may even be comparisons among groups or populations, in accordance with the general specification outlined in the Introduction.

VII. Tests for a single functional relationship

(*a*) *Preliminary remarks*

In § VI it was pointed out that consistent estimates of a functional relationship between two variates could be determined provided there was one instrumental variate. In general, a functional relationship among p variates can be determined if there are $p-1$ or more instrumental variates. In this section we shall consider, not the determination of the functional relationship, but tests for the concordance with the data of a given relationship. This approach parallels that adopted above in the sections on discriminant functions. It is possible to test the adequacy of the given relationship even when the data do not provide enough information to enable consistent estimates to be determined.

We shall consider, following the set-up given in the earlier sections, that there are p investigational variables X_i and q instrumental variables Y_j. The given functional relationship among the X_i will be defined by a null function ξ. It will be realized that, when p exceeds q, it is possible to choose $p-q$ independent linear functions of the original X_i whose correlation with the Y_j (or sum of squares between groups) vanishes. These functions are equivalent to the $p-q$ canonical variates corresponding to the identically vanishing latent roots. Thus it is not possible to determine from the sample a unique null function.

On the other hand, the test for a given null function is always possible. The general test for the adequacy of the function can be considered in two parts: first, a test for the conformity of the function in direction, with $p-1$ degrees of freedom; and secondly, a test for the residual correlation (or reduced variation between groups), with $q-p+1$ degrees of freedom. When $q<p$, no test for residual correlation is possible; the test for direction then has but q degrees of freedom.

(*b*) *One instrumental variate* ($q = 1$)

In order to gain some insight into the problems involved in testing the adequacy of a specified null function, we consider first the simplest case, where there is but one instrumental variate, Y. In this case, the test is one of the direction of the chosen null function. It may be most simply viewed in the following way. The null variate ξ has a sum of squares

for regression on Y which bears to the total sum of squares a ratio λ. It is required to test whether this is large enough to indicate real association. The discriminant function x based on the sample gives a ratio θ, leaving a residual $1-\theta$, with $n-p$ degrees of freedom. Hence the efficient test, in which the residual has been minimized with respect to all the X_i, is given by the following analysis:

	Degrees of freedom	Sum of squares
Additional accounted for by discriminant function x	$p-1$	$\theta-\lambda$
Direction of null variate	1	λ
Residual	$n-p$	$1-\theta$
Total	n	1

so that $$F_{(1,n-p)} = \frac{(n-p)\lambda}{1-\theta}.$$

In terms of determinantal ratios, we have

$$\lambda = b_{\xi\xi}/t_{\xi\xi}, \quad 1-\theta = |W|/|T|,$$

so that $$F_{(1,n-p)} = (n-p)\frac{b_{\xi\xi}}{t_{\xi\xi}}\frac{|T|}{|W|}.$$

This latter representation would give the simplest way of making the test.

(c) *More than one instrumental variate* ($q \geqslant 2$)

The method given above for $q = 1$ cannot be directly adapted to the more general case. The approach that is adopted is as follows. By hypothesis, the null variate ξ is uncorrelated with the instrumental variates, so that its total variation is homogeneous. In any sample, it is possible to choose a set of $p-1$ linearly independent functions of the X_i which are uncorrelated with ξ; these may be taken to define the direction of ξ in the p-space. We shall denote these $p-1$ functions by X_i' ($i = 1, 2, \ldots, p-1$).

Now consider the analysis of any one of the original X_i, or of any linear function of them not linearly dependent on the X_i'. We can carry out an analysis of covariance of this variate with the $p-1$ variates X_i', to determine the reduced sums of squares for regression on the Y_j, for residual variation and for total variation (or between groups, within groups and total in the analysis of variance model). Then, provided $q \geqslant p$, we have the following effects appearing:

Effect	Degrees of freedom
Difference of regressions	$p-1$
Regression on the Y_j, reduced	$q-p+1$
Residual, reduced	$n-p-q+1$
Total, reduced	$n-p+1$

This analysis tests the existence of any correlation of the X_i with the Y_j, apart from that accounted for by the X_i', and hence tests the correlation of the proposed null variate. The second term tests the residual correlation of ξ with the Y_j and the first term the direction; each may be tested against the reduced residual variance.

The analysis and significance tests are equivalent, whichever X-variate (subject to the above-mentioned conditions) be chosen, since the elimination of the X_i' by covariance effectively reduces the number of residual variates to one. In practice it is often simplest to take the X-variate as ξ itself. The fact that the adjustment of ξ by covariance leaves the total sum of squares unaltered does not affect the results; the degrees of freedom for the total line are still reduced to $n-p+1$.

It may be verified that, when $p = 2$, the analysis is exactly the same as that giving the test of significance for the discriminant function X_1'. This is as it should be, since the hypothesis that X_1' is the population discriminant function is equivalent to the hypothesis that the variate orthogonal to it (i.e. ξ) is a null variate.

The analysis has its parallel in the tests developed for this purpose from Fisher's (1938) results by Tintner (1945, 1946), and based on the values of the latent roots. For the existence of a single null function (or linear functional relationship), Tintner takes as his test function the smallest of p latent roots. As has been shown by Hsu (1941), this smallest latent root is distributed approximately as a sum of squares with $q-p+1$ degrees of freedom. This approximation becomes exact in the limit when all the other latent roots are near to unity (apparently even if only the sample latent roots approach unity, regardless of the values of the population roots). This result has been established above (10) as a particular result on the factorization of determinantal ratios, except that there the term for 'difference of regressions' has not been segregated from the residual. In the derivation of the tests to be given below, in terms of canonical correlations, the relationship of the present tests with the approximate tests will be made clear.

(i) *Analysis in terms of canonical correlations*

The results of the analysis of covariance may be expressed in terms of the squared canonical correlation coefficients or discriminant ratios $\theta_1, \theta_2, \ldots, \theta_p$ between the sets of canonical variates x_i and y_i. Let the hypothetical null-variate be

$$\xi = \Sigma m_i x_i, \quad \Sigma m_i^2 = 1.$$

Then the set of $p-1$ x-variates orthogonal to ξ may be taken as

$$x_i' = x_i - m_i \xi.$$

Now of the y_i, p have non-zero correlation with the corresponding x_i, and $q-p$ have zero correlation with any x_i. They may be transformed to another set of variates in the following way: (1) the set of $p-1$ variates most closely correlated with the x_i'; (2) a variate η, uncorrelated with the x_i' but correlated with the x_i, and (3), as before, the $q-p$ variates uncorrelated with the x_i.

It is clear that the variate η represents the reduced correlation of ξ with the y_i. It is readily defined by the conditions given; for let

$$\eta = \Sigma n_i y_i;$$

then

$$\Sigma \eta x_h' = 0 \quad (h = 1, 2, \ldots, p-1).$$

Hence $$\Sigma\eta x_h = m_h \Sigma\xi\eta,$$

or $$n_h\theta_h^{\frac{1}{2}} = m_h \Sigma m_i n_i \theta_i^{\frac{1}{2}},$$

so that $$n_h \propto m_h \theta_h^{-\frac{1}{2}},$$

and, since $$\Sigma n_h^2 = 1,$$

$$n_h = m_h \theta_h^{-\frac{1}{2}} (\Sigma m_i^2/\theta_i)^{-\frac{1}{2}}.$$

For the regression of ξ on η, the sum of squares, with $q-p+1$ degrees of freedom, is

$$(\Sigma\xi\eta)^2 = n_h^2 \theta_h / m_h^2 = (\Sigma m_i^2/\theta_i)^{-1}. \tag{18}$$

This is the sum of squares for regression, reduced by elimination of the x_i'. Clearly it vanishes if $q < p$ (for then $\theta_p = 0$), and in general takes values between θ_p and θ_1.

In the same way it may be shown that the reduced residual sum of squares (i.e. the sum of squares of residuals of ξ from regression on the y_i and the x_i'), with $n-p-q+1$ degrees of freedom, is

$$(\Sigma m_i^2/(1-\theta_i))^{-1}. \tag{19}$$

Finally, by subtraction, the sum of squares for difference of regressions (i.e. for the test of direction of the proposed null-variate), with $p-1$ degrees of freedom, is

$$1-(\Sigma m_i^2/\theta_i)^{-1} - (\Sigma m_i^2/(1-\theta_i))^{-1}. \tag{20}$$

These sums of squares may also be expressed in terms of the latent roots θ_i and the latent roots for the variates x_i' and y_i, which will be denoted by θ_i' ($i = 1, 2, \ldots, p-1$). Since ξ is orthogonal to all the x_i', it follows that the product of the latent roots θ_i', multiplied by the sum of squares for regression of ξ on η, is equal to the product of the original latent roots; that is,

$$\Pi\theta_i'(\Sigma m_i^2/\theta_i)^{-1} = \Pi\theta_i.$$

Similarly, $$\Pi(1-\theta_i')(\Sigma m_i^2/(1-\theta_i))^{-1} = \Pi(1-\theta_i).$$

The analysis of variance may then be set out as follows:

	Degrees of freedom	Sum of squares
Direction	$p-1$	$1-\Pi\theta_i/\Pi\theta_i' - \Pi(1-\theta_i)/\Pi(1-\theta_i')$
Correlation	$q-p+1$	$\Pi\theta_i/\Pi\theta_i'$
Residual	$n-p-q+1$	$\Pi(1-\theta_i)/\Pi(1-\theta_i')$
Total	$n-p+1$	1

If the latent roots θ_i and the direction cosines m_i are known, the simplest procedure is to calculate the sums of squares from the expressions (18), (19) and (20).

The sum of squares for correlation always exceeds θ_p, but approaches θ_p when the direction of ξ approaches that of x_p. Hence it is seen that the use of θ_p as a sum of squares with $q-p+1$ degrees of freedom, according to the approximate tests, will underestimate the significance of the correlation.

If $q < p$, the analysis reduces to

	Degrees of freedom	Sum of squares
Direction	q	$1-\Pi(1-\theta_i)/\Pi(1-\theta'_i)$
Residual	$n-p-q+1$	$\Pi(1-\theta_i)/\Pi(1-\theta'_i)$
Total	$n-p+1$	1

there being in this case no test for correlation.

When $q = 1$ this analysis reduces to that previously given; for it may be shown that, in general,

$$\lambda = \Sigma\theta_i - \Sigma\theta'_i,$$

so that, in this case $$(1-\theta_1)/(1-\theta'_1) = 1-\lambda/(1-\theta_1+\lambda)$$

and the sums of squares for direction and residual are respectively

$$\lambda/(1-\theta_1+\lambda)$$

and $$(1-\theta_1)/(1-\theta_1+\lambda),$$

which are proportional to those previously given.

It may readily be verified that, when $p = 2$, these sums of squares reduce to those given for a test of a single discriminant function.

(ii) *Analysis in terms of determinantal ratios*

For computing purposes it is more straightforward, when the individual latent roots and canonical variates are not required, to express the results just given in terms of determinantal ratios.

We consider the analysis of the null variate ξ ($= \Sigma l_i X_i$, no longer necessarily normalized so that its sum of squares is unity), deriving the reduced sum of squares for regression, with $q-p+1$ degrees of freedom, and the reduced residual sum of squares, with $n-p-q+1$ degrees of freedom. The sum of squares for 'difference of regressions', with $p-1$ degrees of freedom, is obtained by subtraction of these two items from the (reduced) total.

We may take the X'_i to be

$$X'_i = X_i - \xi t_{\xi i}/t_{\xi\xi} \quad (i = 1, 2, \ldots, p-1),$$

and shall write

P_i for the vector $(p_{i1}, p_{i2}, \ldots, p_{iq})$,
T_i for the vector $(t_{i1}, t_{i2}, \ldots, t_{ip})$,
T_{i-} for the vector $(t_{i1}, t_{i2}, \ldots, t_{i,p-1})$,
P_- for the matrix consisting of the first $p-1$ rows of P,
and T_- for the matrix consisting of the first $p-1$ rows and columns of T.

To calculate the reduced sum of squares for regression, the sums of squares and products of the X'_i and ξ for the regression of each on the Y_j are first determined. For example, the sum of products of X'_i with Y_j is

$$p_{ij} - p_{\xi j} t_{\xi i}/t_{\xi\xi},$$

so that the regression sum of products of X'_h and X'_i is

$$(P_h - P_\xi t_{\xi h}/t_{\xi\xi})\, U^{-1}\, (P'_i - P'_\xi\, t_{\xi i}/t_{\xi\xi}).$$

Accordingly, the matrix of regression sums of squares and products of the X'_i is

$$(P_- - T'_{\xi-} P_\xi/t_{\xi\xi})\, U^{-1} (P'_- - P'_\xi T_{\xi-}/t_{\xi\xi}). \tag{21}$$

Now the reduced sum of squares for the regression of ξ on the Y_j is the ratio of two determinants: that of the regression sums of squares and products of the X'_i and ξ, and that of the regression sums of squares and products of the X'_i alone. By transformation without change of scale, the X'_i and ξ may be replaced by $X_1, X_2, \ldots, X_{p-1}$, and $l_p X_p$; hence the determinant in the numerator is equal to

$$l_p^2\, |\, PU^{-1}P'\, | = l_p^2\, |\, B\, |.$$

The required denominator is the determinant of the matrix (21), which is found, by elementary transformations, to be

$$-\frac{l_p^2}{t_{\xi\xi}^2} \begin{vmatrix} 0 & T_\xi \\ T'_\xi & PU^{-1}P' \end{vmatrix} = -\frac{l_p^2}{t_{\xi\xi}^2} \begin{vmatrix} 0 & T \\ T' & B \end{vmatrix}$$

$$= \frac{l_p^2}{t_{\xi\xi}^2}\, |\, B\, |\, T_\xi B^{-1} T'_\xi.$$

Hence the reduced sum of squares for regression is

$$t_{\xi\xi}^2/T_\xi B^{-1} T'_\xi.$$

In the same way, it may be shown that the reduced residual sum of squares is

$$t_{\xi\xi}^2/T_\xi W^{-1} T'_\xi.$$

Also, since

$$T_\xi T^{-1} T'_\xi = t_{\xi\xi},$$

we see that the reduced total sum of squares, which is given by the analogous formula

$$t_{\xi\xi}^2/T_\xi T^{-1} T'_\xi$$

is equal to $t_{\xi\xi}$, as it should be.

On elimination of common factors from each term, the analysis of variance may be set out as follows:

	Degrees of freedom	Sum of squares
Direction	$p-1$	(by subtraction)
Regression, reduced	$q-p+1$	$(T_\xi B^{-1} T'_\xi)^{-1}$
Residual, reduced	$n-p-q+1$	$(T_\xi W^{-1} T'_\xi)^{-1}$
Total, reduced	$n-p+1$	$(T_\xi T^{-1} T'_\xi)^{-1} = t_{\xi\xi}^{-1}$

It may readily be verified that, when the variates are transformed to the canonical set, the terms in this analysis of variance reduce to those given in the previous subsection.

VIII. Generalizations

It has been shown above that the test for a given functional relationship as developed in this paper is really equivalent to the test for the adequacy of $p-1$ discriminant functions to represent the variation in the data. Thus we have treated here only the tests for 1 or $p-1$ discriminant functions. Tests for any number r of discriminant functions may be derived in the same way, but the cases treated above are the simplest to deal with, since they each deal with only one function (discriminant function or null function); fiducial limits for the constants in these functions can therefore be derived using these significance tests.

If r hypothetical discriminant functions are given, their effect may be eliminated, giving the following factorization of the likelihood criterion:

$$(n; p, q) = (n; r, q)\,(n-r; p-r, q).$$

The second factor on the right is the residual likelihood criterion, which may be further factorized to give criteria for direction and collinearity:

$$\begin{aligned}(n-r; p-r, q) &= \underset{\text{simple direction}}{(n-r; p-r, r)} \quad \underset{\text{partial collinearity}}{(n-2r; p-r, q-r)} \\ &= \underset{\text{partial direction}}{(n-q; p-r, r)} \quad \underset{\text{simple collinearity}}{(n-r; p-r, q-r)}\end{aligned}$$

When $r = p-1$, one of the degrees of freedom in each factor is unity, which is the reason why the tests for a single null function can be thrown into the form of an analysis of variance. We then have

$$\begin{aligned}(n-p+1; 1, q) &= (n-p+1; 1, p-1) \quad (n-2p+2; 1, q-p+1) \\ &= (n-q; 1, p-1) \quad (n-p+1; 1, q-p+1)\end{aligned}$$

which corresponds to the analysis of variance set-up previously given.

There seem to be practical applications of tests for more than one functional relationship; for example, Bartlett (1948) discusses the problem of estimating supply and demand relationships. On the other hand, it seems that seldom if ever is more than one discriminant function required in specifying differences among populations; if these differences cannot be interpreted in terms of one discriminant function, what is then required is some likelihood test as a means of classifying the observations into the different populations.

REFERENCES

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* **22**, 327.

Bartlett, M. S. (1938). Further aspects of the theory of multiple regression. *Proc. Camb. Phil. Soc.* **34**, 33.

Bartlett, M. S. (1939). A note on tests of significance in multivariate analysis. *Proc. Camb. Phil. Soc.* **35**, 180.

Bartlett, M. S. (1947). Multivariate analysis. *J. R. Statist. Soc. Supp.* **9**, 176.

Bartlett, M. S. (1948). A note on the statistical estimation of supply and demand relations from time series. *Econometrica*, **16**, 323.

Bartlett, M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics*, **5**, 207.

Bartlett, M. S. (1951). The goodness of fit of a single hypothetical discriminant function in the case of several groups. *Ann. Eugen., Lond.*, **16**, 199.

Berkson, J. (1950). Are there two regressions? *J. Amer. Statist. Ass.* **45**, 164.

FISHER, R. A. (1938). The statistical utilization of multiple measurements. *Ann. Eugen., Lond.*, **8**, 376.
GEARY, R. C. (1943). Relations between statistics: the general and the sampling problem when the samples are large. *Proc. R. Irish Acad.* A, **49**, 177.
GEARY, R. C. (1948). Studies in relations between economic time series. *J. R. Statist. Soc.* B, **10**, 140.
GEARY, R. C. (1949). Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown. *Econometrica*, **17**, 30.
HAAVELMO, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, **11**, 1.
HSU, P. L. (1941). On the problem of rank and the limiting distribution of Fisher's test function. *Ann. Eugen., Lond.*, **11**, 39.
KOOPMANS, T. (1937). *Linear Regression Analysis of Economic Time Series.* Haarlem: DeErven, F., Bohn, N.V.
LINDLEY, D. V. (1947). Regression lines and the linear functional relationship. *J. R. Statist. Soc. Supp.* **9**, 218.
LINDLEY, D. V. (1953). Estimation of a functional relationship. *Biometrika*, **40**, 47.
NEYMAN, J. & SCOTT, ELIZABETH L. (1951). On certain methods of estimating the linear structural relation. *Ann. Math. Statist.* **22**, 352.
REIERSØL, O. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica*, **9**, 1.
REIERSØL, O. (1945). Confluence analysis by means of instrumental sets of variables. *Ark. Mat. Astr. Fys.* A, **32**, no. 4.
REIERSØL, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica*, **18**, 375.
TINTNER, G. (1945). A note on rank, multicollinearity and multiple regression. *Ann. Math. Statist.* **16**, 304.
TINTNER, G. (1946). Multiple regression for systems of equations. *Econometrica*, **14**, 5.
TINTNER, G. (1950). A test for linear relations between weighted regression coefficients. *J. R. Statist. Soc.* B, **12**, 273.
WALD, A. (1940). The fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.* **11**, 284.
WILLIAMS, E. J. (1952*a*). Some exact tests in multivariate analysis. *Biometrika*, **39**, 17.
WILLIAMS, E. J. (1952*b*). The interpretation of interactions in factorial experiments. *Biometrika*, **39**, 65.
WILLIAMS, E. J. (1952*c*). Use of scores for the analysis of association in contingency tables. *Biometrika*, **39**, 274.
WILLIAMS, E. J. (1953). Tests of significance for concurrent regression lines. *Biometrika*, **40**, 297.

THE USE OF TRANSFORMATIONS AND MAXIMUM LIKELIHOOD IN THE ANALYSIS OF QUANTAL EXPERIMENTS INVOLVING TWO TREATMENTS

By F. YATES, F.R.S.

Rothamsted Experimental Station

If the effect of a treatment is of the quantal (all or nothing) type, e.g. it kills or does not kill an insect, the results of experiments will be in the form of proportions of experimental units affected. If t such treatments are compared in a single experiment in which the experimental units are randomly allocated, the differences between the observed proportions will give estimates of the differences between the treatments, and a χ^2 test on the associated $2 \times t$ table ($t-1$ d.f.) provides an overall test of significance. If, however, in order to increase the precision, the units are subdivided before allocation into groups which are relatively homogeneous within themselves, the experiment will be analogous to an ordinary randomized block experiment, the groups being the blocks, but the quantities requiring analysis (corresponding to the 'plot yields') will be proportions instead of quantitative measurements. When a set of experiments is carried out, with random allocation within each experiment, but with possible lack of homogeneity between experiments, the situation is similar, the groups being in this case experiments.

If the numbers of experimental units in the various cells of the group × treatment or experiment × treatment table are approximately equal, or more generally if they are proportionate, the analysis presents no great difficulty. The use of the angular transformation, and an analysis of variance of the transformed values, will frequently be all that is required. Indeed, if the variation in susceptibility from experiment to experiment is not large, the results will not be seriously distorted if the data from the different groups or experiments are pooled and treated as homogeneous. If, however, the numbers in the different cells vary in an irregular manner, as is often the case in material of this kind, pooling is inadmissible and inequality in the weights prevents straightforward analysis.

The application of methods of estimation to quantal data has been extensively developed in recent years through the use of transformations in conjunction with maximum likelihood. Exact analytical methods were first developed for the probit transformation, in connexion with biological assay, but similar methods have recently been developed for other transformations of which the logit and log log are of most interest in this connexion. These methods are beginning to be used for the analysis of quantal data of many kinds (see, for example, Jolly, 1950; Dyke & Patterson, 1952; and Yates, 1953, § 9·7), but their utility for dealing with experimental data of the type we are considering is still insufficiently realized. Combination of probabilities, for example, is still commonly used to provide a test of significance (sometimes called the probability integral test) for sets of experiments involving two treatments.

The estimation approach automatically provides a test of significance as a by-product of the estimate and its estimate of error. Its greatest advantage, however, is that if an effect is demonstrated we are immediately provided with an estimate of its magnitude in

meaningful terms, together with approximate fiducial limits. At the same time, answers are provided to certain ancillary questions regarding the homogeneity of the data.

The analysis of a group of experiments containing several treatments has been discussed by Jolly using the probit transformation. The present paper deals in more detail with the case of two treatments, particularly in respect of errors and tests of significance. The appropriateness of the various transformations under different circumstances is also considered in more detail than by Jolly. A further new point is the use of weighted estimates to give a first approximation. The investigation originated in a request by Prof. Gert Bonnier of Stockholm for advice on the exact procedure for applying the combination of probabilities test to a set of genetical experiments on mutation rates in *Drosophila*, and I am much indebted to him for permission to reproduce the data here. The numbers involved in these experiments were sufficiently large for the data to be analysed by large-sample methods, and this analysis will first be described, as it brings out many of the points at issue without the complexities associated with the use of transformations and the method of maximum likelihood. The maximum-likelihood method will then be developed, and applied to this set of data and to three examples given by Pearson (1950).

THE COMBINATION OF PROBABILITIES TEST

If we have a set of k experiments (or observational data) comparing two treatments, the results of each experiment can be set out in the form of a 2×2 table. A test of significance on the data as a whole can be made by calculating the significance level P for each experiment separately, and then combining these probabilities by the method first suggested by Fisher in the fourth edition of *Statistical Methods for Research Workers* (1932), and independently by Karl Pearson (1933). This consists of calculating the sum of the quantities $-2\log_e P$, which are the values of χ^2 for 2 d.f. corresponding to the significance levels P. If there is no difference between the treatments, then, subject to certain qualifications, $S(-2\log_e P)$ will be distributed as χ^2 with $2k$ d.f.

This test and the analogous test provided by direct summation of the values of χ^2 for the separate experiments have been discussed by various authors, in particular Cochran (1942, 1952), Lancaster (1949) and Pearson (1950). Certain further points emerged in the course of the present investigation, but these are best dealt with in a separate paper (Yates, 1955). Here it is only necessary to make the following general points:

(1) Application of the combination of probabilities test does not provide an analysis which is in any sense complete, since no estimate of the magnitude of the treatment difference is provided.

(2) The test must be in some degree inefficient. There are two main reasons for this:

(*a*) No account is taken of the relative accuracy of the different experiments when combining the results.

(*b*) $S(-2\log_e P)$ does not provide an efficient estimate of the treatment difference, even when all experiments are of the same accuracy and the difference is small, and consequently cannot be the basis of an efficient test of significance.

It is perhaps also worth noting here that Fisher did not envisage the use of the combination of probabilities test on data of the type we are considering. He put it forward for use in cases in which it is desired to obtain a single test of the significance of an aggregate of probabilities, 'taking account only of these probabilities, and not of the detailed composition of the data from which they are derived, which may be of very different kinds'.

Drosophila data: the large sample approach

The data with which we are concerned are shown in Table 1. They were derived from a set of six experiments on the frequency of lethals in paternal *X*-chromosomes from irradiated spermatozoa of *Drosophila melanogaster*. The experiments were carried out to test whether there was any difference in this frequency when the irradiation was given to spermatozoa which were harboured (F) in the females' receptacles, and (M) in the males' testicles. In the first four experiments a dose of 960 r. was given, and in the last two a dose of 3000 r. In addition to the change of irradiation rate it was believed that unavoidable environmental changes between experiments might influence the mutation rates. An account of the experiments has now been published (Bonnier & Lüning, 1953).

Table 1. *Frequency of mutants with two methods of irradiation*

Exp.	Method F			Method M		
	Mutant	Normal	% mutant	Mutant	Normal	% mutant
1	29	815	3·44	45	1486	2·94
2	78	2622	2·89	27	1438	1·84
3	110	3175	3·35	100	4281	2·28
4	25	1038	2·35	52	2053	2·47
5	31	339	8·38	43	507	7·82
6	57	543	9·50	27	436	5·83

It will be seen that in all experiments except Exp. 4 method F gives a higher mutation rate than method M. We require to test the significance of this difference, and obtain an estimate of its magnitude.

In order to develop a suitable estimation procedure we must specify in mathematical terms the quantities that require estimation. This specification will depend on the phenomena that are being investigated. In the present case we may postulate that with a given method of irradiation and an experiment of standard sensitivity a unit of irradiation has a given small chance θ' of producing a mutation at a given locus. The chance that λ units of irradiation will produce a mutation at that locus will then be

$$1-e^{-\lambda\theta'}.$$

If several loci are involved, with probabilities θ', θ'', etc., and $\theta'+\theta''+\ldots=\theta$, the total probability of one or more mutations will be

$$1-e^{-\lambda(\theta'+\theta''+\ldots)}=1-e^{-\lambda\theta}.$$

If the conditions vary from experiment to experiment a sensitivity factor μ_r will have to be introduced, giving a total probability of mutation for experiment r of $1-e^{-\lambda_r\mu_r\theta}$, where λ_r is the irradiation rate for the rth experiment.

With two methods of irradiation the probabilities of mutation in experiment r can therefore be written as

$$\pi_{1r}=1-\exp[-\lambda_r\mu_r\theta_1],$$
$$\pi_{2r}=1-\exp[-\lambda_r\mu_r\theta_2].$$

Since μ_r is unknown it must be eliminated, giving

$$\frac{\theta_1}{\theta_2} = \frac{\log(1-\pi_{1r})}{\log(1-\pi_{2r})}.$$

Substituting the observed mutation rates p_{1r} and p_{2r} for π_{1r} and π_{2r} we obtain an estimate of θ_1/θ_2 from the rth experiment. These separate estimates can then be combined by weighting inversely as their estimated variances.

In the present example the mutation rates are sufficiently small for π_{1r} and π_{2r} to be taken as approximately equal to $\lambda_r\mu_r\theta_1$, and $\lambda_r\mu_r\theta_2$ respectively. There will be certain advantages in working with estimates of

$$\frac{\theta_1-\theta_2}{\frac{1}{2}(\theta_1+\theta_2)}$$

instead of θ_1/θ_2, i.e. the ratio of the difference of the mutation rates for experiments of unit sensitivity per unit irradiation to the mean of these mutation rates. The quantity

$$E_r = \frac{p_{1r}-p_{2r}}{\frac{1}{2}(p_{1r}+p_{2r})}$$

will provide an estimate of this ratio. An approximation to the variance of E_r is easily calculated by large-sample theory as follows:

$$\frac{\partial E_r}{\partial p_{1r}} = \frac{4p_{2r}}{(p_{1r}+p_{2r})^2}, \quad \frac{\partial E_r}{\partial p_{2r}} = \frac{-4p_{1r}}{(p_{1r}+p_{2r})^2}.$$

If the rth contingency table is written

	F	M	Total
Mutant	n'_{1r}	n'_{2r}	n'_r
Normal	$n_{1r}-n'_{1r}$	$n_{2r}-n'_{2r}$	$n_r-n'_r$
Total	n_{1r}	n_{2r}	n_r

so that $p_{1r} = n'_{1r}/n_{1r}$, $p_{2r} = n'_{2r}/n_{2r}$, we have

$$V(p_{1r}) = \frac{\pi_{1r}(1-\pi_{1r})}{n_{1r}}, \quad V(p_{2r}) = \frac{\pi_{2r}(1-\pi_{2r})}{n_{2r}}, \quad \operatorname{cov} p_{1r}p_{2r} = 0.$$

Hence, approximately,

$$V(E_r) = \frac{16p_{1r}p_{2r}}{(p_{1r}+p_{2r})^4}\left[\frac{p_{2r}q_{1r}}{n_{1r}} + \frac{p_{1r}q_{2r}}{n_{2r}}\right]. \tag{1}$$

This formula is approximate both because we are dealing with a non-linear function and because π_{1r} and π_{2r} have been replaced by p_{1r} and p_{2r}. If we are concerned with testing the significance of the difference of p_{1r} and p_{2r} we may replace them both by the pooled estimate $p_r = n'_r/n_r$ from the marginal totals of the table. Since the factor $1/(p_{1r}+p_{2r})$ occurs in E_r the factor $1/(p_{1r}+p_{2r})^2$ may be left unchanged in the variance. We then have

$$V(E_r) = \frac{4}{(p_{1r}+p_{2r})^2}\frac{p_rq_rn_r}{n_{1r}n_{2r}}. \tag{2}$$

It may be noted that under these circumstances

$$\frac{E_r^2}{V(E_r)} = \frac{n_r\{n_{1r}(n_{2r}-n'_{2r})-n_{2r}(n_{1r}-n'_{1r})\}^2}{n_{1r}n_{2r}n'_r(n_r-n'_r)},$$

which is the ordinary expression for χ^2 in a 2×2 contingency table, without correction for continuity.

As has been shown by Cochran (1942) and others, the correction for continuity is not required when combining data from several tables. If, therefore, we take a weighted mean of the estimates E_r with weights $1/V(E_r)$, the resultant test of significance may be expected to be adequately accurate when the values of χ^2 corrected for continuity give satisfactory approximations to the one-tail probabilities of the distributions generated by the 2×2 tables. This will be the case in the present instance, since all expectations are reasonably large and n_{1r} and n_{2r} differ by less than a factor of 2 in all experiments, and are very nearly equal in total.

The calculations for $\bar{E}$ using formula (2) for $V(E_r)$ are shown in Table 2. The weighted mean of E is

$$\bar{E} = 44{\cdot}182/153{\cdot}85 = 0{\cdot}2872,$$

with standard error $1/\sqrt{153{\cdot}85} = 0{\cdot}0806$. The level of significance (one tail) is therefore $P = 0{\cdot}000184$.

Table 2. *Calculation of* $\bar{E}$

Exp.	E	$V(E)$	w	wE
1	+0·1559	0·05459	18·32	2·856
2	+0·4421	0·04622	21·63	9·563
3	+0·3786	0·01790	55·88	21·156
4	−0·0489	0·05776	17·31	−0·846
5	+0·0692	0·05099	19·61	1·357
6	+0·4785	0·04739	21·10	10·096
	+0·2872 ± 0·0806	—	153·85	44·182

$P = 0{\cdot}000184$.

The results of Table 2 also provide a simple test of the constancy of $(\theta_1-\theta_2)/\frac{1}{2}(\theta_1+\theta_2)$ over the different experiments. All that is necessary is to calculate the weighted sum of squares of deviations of E. This is

$$\begin{aligned} Sw_r(E_r-\bar{E})^2 &= Sw_rE_r^2-\bar{E}Sw_rE_r \\ &= 4{\cdot}962. \end{aligned}$$

If there is no real variation this will be distributed as χ^2 with 5 d.f. This gives $P = 0{\cdot}42$, so there is no evidence of variation.

The same procedure may be adopted to test whether there is any evidence of variation in the mutation rates from experiment to experiment, apart from those due to differences in irradiation rate. For this purpose it will be appropriate to test whether the estimates of

$\frac{1}{2}(\theta_1+\theta_2)$ are homogeneous. These estimates are given in Table 3, together with their standard errors calculated in the conventional manner. The value of the weighted sum of squares of deviations of these rates is 5·32 (χ^2 with 5 d.f.) giving $P = 0{\cdot}38$. There is thus no evidence of any such variation. The postulated law also accounts satisfactorily for the difference in mutation rates at the different levels of irradiation.

Table 3. *Estimates of* $\frac{1}{2}(\theta_1+\theta_2) \times 10^4$

Exp.	
1	0·332 ± 0·0396
2	0·246 ± 0·0248
3	0·293 ± 0·0201
4	0·251 ± 0·0300
5	0·270 ± 0·0306
6	0·256 ± 0·0270
Weighted mean	0·272 ± 0·0110

It may be noted that if the irradiation rate had been constant, and the difference between θ_1 and θ_2 could be neglected, the ordinary χ^2 test on the 2×6 contingency table of the values of n_r' and $n_r - n_r'$ could be used to test the constancy of the mutation rate from experiment to experiment.

Since the difference between the p_{1r} and p_{2r} is demonstrated we might revise the calculations by using formula (1) for $V(E_r)$. This, however, is merely a partial step in the direction of the small sample solution by maximum likelihood, which will be outlined in the following sections.

Cochran (1954) has adopted a similar procedure to that given above in order to provide a test of significance for a set of 2×2 contingency tables. His test criterion is the weighted mean of the quantities $p_{1r} - p_{2r}$. This is, in fact, equivalent to a weighted mean of the quantities

$$E_r' = \frac{p_{1r} - p_{2r}}{p_r q_r}.$$

E_r' differs from E_r only in the substitution of the pooled estimate p_r and the introduction of the factor q_r, which gives a difference, analogous to logits, which may be expected to be reasonably constant over a wide range of p_r. Regarded in this way, therefore, Cochran's analysis could easily be extended to provide a test for the constancy of the difference from experiment to experiment.

MAXIMUM-LIKELIHOOD SOLUTION: CHOICE OF TRANSFORMATION

The maximum-likelihood solution is most easily obtained by the use of a suitable transformation, as, for example, the probit transformation, which has become familiar in toxicology and biological assay. The maximum-likelihood solution can then be obtained by successive approximation. The choice of transformation depends on the mathematical specification (the 'model') which is considered most appropriate to the data under

investigation. When this specification has been decided the transformation chosen should be such that the quantities that require estimation are simple functions of the transformed variate; thus in probit work it is anticipated that the transformed variate will bear a linear relation to the dosage when the latter is expressed in suitable units (usually logarithms), and the quantities that require estimation are the constant term and the slope of this regression. A secondary requirement is that in no part of the range (or at least in no part covered by the data) do the working values or the weights associated with the transformed variate become infinite.

In the genetical example given above the transformation

$$y = \log(-\log q)$$

is obviously suitable. For with this transformation if $\mathbf{Y}_{1r}$ and $\mathbf{Y}_{2r}$ represent the transformed values of π_{1r} and π_{2r} we have

$$\mathbf{Y}_{1r} = \log\lambda_r + \log\mu_r + \log\theta_1,$$

$$\mathbf{Y}_{2r} = \log\lambda_r + \log\mu_r + \log\theta_2.$$

The effects of the treatments and of the variations in sensitivity are therefore represented by additive components in the transformed scale. This transformation, which is the same as the ordinary log log transformation with p replaced by q, may be termed the *complementary log log transformation*.

In many cases in which there is no clearly appropriate theoretical model, and in which there is no reason to differentiate between the two ends of the probability scale, the logit transformation

$$z = \tfrac{1}{2}\log(p/q)$$

is likely to be suitable. When the p's are all small the logit transformation is equivalent to one-half the complementary log log transformation, and when the p's are all nearly unity the logit transformation is equivalent to one-half the ordinary log log transformation.*

If the logit transformation is used the analysis will be conducted on the assumption that the treatment and sensitivity effects are additive on the transformed scale, i.e. that, except for random variations, the difference in the logits for the two treatments is constant from experiment to experiment.

In certain circumstances the probit transformation may be considered to be more appropriate than the logit transformation. Except at the extreme ends of the scale, however, there is little difference between them, and in the absence of any theoretical reasons for choosing one or the other it will rarely be possible to determine, by reference to the data themselves, which specification gives a better representation of the effects under investigation.

Maximum-likelihood equations

As has been demonstrated in a number of places, the general procedure of obtaining the maximum-likelihood solution when using a transformation of the form $y = \phi(p)$, is as follows:

(1) Determine a set of N provisional values Y of the transformed variate for the N observed proportions. These can conveniently be deduced from preliminary (here called first) estimates of the required parameters or, when regression lines are involved, graphically.

* In the author's *Sampling Methods* (2nd edition, 1953) the logit transformation is defined as $z = \log p/q$. This does not accord with modern usage, and was due to an over-hasty reference to Finney's *Statistical Method in Biological Assay* (1952).

In many cases it pays, when making the first estimate, to use weights similar to those used in the further steps of the analysis.

(2) For each observed proportion p (given by say n' successes out of n observations) calculate a working value given (when y increases uniformly with p) by

$$\begin{aligned} y &= pY_{\text{max.}} + qY_{\text{min.}} \\ &= Y_{\text{max.}} - q(\text{Range}) \\ &= Y_{\text{min.}} + p(\text{Range}). \end{aligned}$$

where $Y_{\text{max.}}$ and $Y_{\text{min.}}$ are the maximum and minimum working values corresponding to the assigned provisional value Y, and $Y_{\text{max.}} - Y_{\text{min.}} = \text{Range}$. Determine also the weight W to be assigned to the working value. This is given by nw, where w is the weighting coefficient. $Y_{\text{max.}}$, $Y_{\text{min.}}$ and w are functions of Y (or of the corresponding untransformed proportion P) whose form depends on the transformation. The general formulae are

$$Y_{\text{max.}} = Y + Q\left(\frac{dY}{dP}\right), \quad Y_{\text{min.}} = Y - P\left(\frac{dY}{dP}\right), \quad \text{Range} = \left(\frac{dY}{dP}\right), \quad w = \frac{1}{PQ}\Big/\left(\frac{dY}{dP}\right)^2.$$

For the transformations commonly in use tables of these functions are readily available.

If y decreases uniformly with increasing p the formulae for $Y_{\text{max.}}$ and $Y_{\text{min.}}$ are interchanged and p and q are interchanged in the formulae for the working value y.

(3) Using the working values as observed variates obtain revised estimates of the parameters by the ordinary method of least squares, weighting each observation by the weight determined as above.

(4) From the revised estimates of the parameters determine a new set of provisional values, and repeat the whole process as often as is necessary.

In the case of the genetic example already discussed in large sample terms we may write

$$\begin{aligned} &\log\theta_1 = t_1, \quad \log\theta_2 = t_2, \\ &\log\lambda_r = l_r, \quad \log\mu_r = m_r \quad (r = 1, \ldots, k). \end{aligned}$$

t_1, t_2 and the m_r are the parameters that have to be estimated.

If t_1', t_2' and m_r' are first estimates of these parameters provisional values are given by the equations

$$\left.\begin{aligned} Y_{1r} &= l_r + m_r' + t_1', \\ Y_{2r} &= l_r + m_r' + t_2'. \end{aligned}\right\} \tag{3}$$

Experiment r gives the two observation equations

$$\begin{aligned} t_1 + m_r &= y_{1r} - l_r + \epsilon_{1r} \text{ weight } W_{1r}, \\ t_2 + m_r &= y_{2r} - l_r + \epsilon_{2r} \text{ weight } W_{2r}, \end{aligned}$$

there being k such pairs of equations in all.

The ordinary method of least squares then gives the normal equations:

$$\begin{aligned} t_1\Sigma W_{1r} + \Sigma m_r W_{1r} &= \Sigma W_{1r}(y_{1r} - l_r), \\ t_2\Sigma W_{2r} + \Sigma m_r W_{2r} &= \Sigma W_{2r}(y_{2r} - l_r); \\ m_r(W_{1r} + W_{2r}) + t_1 W_{1r} + t_2 W_{2r} &= W_{1r}(y_{1r} - l_r) + W_{2r}(y_{2r} - l_r) \quad (r = 1, \ldots, k), \end{aligned}$$

where Σ denotes summation over r from 1 to k and t_1, t_2 and the m_r now denote estimates. These equations are of the same form as equation (1) of Jolly's paper.

Since one of these equations is redundant we may conveniently put

$$\Sigma m_r(W_{1r}+W_{2r}) = 0. \tag{4}$$

The sum of the first two equations then gives

$$t_1\Sigma W_{1r}+t_2\Sigma W_{2r} = \Sigma W_{1r}y_{1r}+\Sigma W_{2r}y_{2r}-\Sigma(W_{1r}+W_{2r})l_r. \tag{5}$$

Substituting the values of m_r given by the last k equations in either of the first two equations, we obtain

$$t_1-t_2 = \left\{\Sigma\frac{W_{1r}W_{2r}}{W_{1r}+W_{2r}}(y_{1r}-y_{2r})\right\}\Big/\Sigma\frac{W_{1r}W_{2r}}{W_{1r}+W_{2r}}. \tag{6}$$

Thus t_1-t_2 is a weighted mean of $y_{1r}-y_{2r}$, with weights $W_{1r}W_{2r}/(W_{1r}+W_{2r})$.

To obtain m_r equations (5) and (6) may be solved for t_1 and t_2. We then have

$$m_r+l_r = \frac{1}{W_{1r}+W_{2r}}\{W_{1r}(y_{1r}-t_1)+W_{2r}(y_{2r}-t_2)\}. \tag{7}$$

When the logit transformation is used and the treatment and sensitivity effects are assumed additive on the logit scale, the solution is identical with that given above, except for the omission of the l terms.

Application to *Drosophila* data

The computations leading to the normal equations are set out in Table 4. The first estimates have been obtained by taking values of y corresponding to the observed p, and weighting coefficients corresponding to these y. Thus from Table XVI of Finney (1952) the y value corresponding to $1-0{\cdot}0344$ is $-3{\cdot}37$, and from Table XVII the corresponding weighting coefficient is $0{\cdot}033$, so that $W = 0{\cdot}033\times 844 = 28$. The values of y_1-y_2, W_1+W_2 and $W_1W_2/(W_1+W_2)$ are then calculated. These last are used as weights in calculating the weighted mean of y_1-y_2, namely, $+0{\cdot}2990$. The equations for t_1' and t_2' are therefore

$$332t_1'+290t_2' = -6505{\cdot}33,$$
$$t_1'-t_2' = +0{\cdot}2990.$$

The numerical term of the first equation is obtained from the sum of products of the y and W columns less those of the l and W_1+W_2 columns.

Solution in the ordinary manner gives

$$t_1' = -10{\cdot}3193, \quad t_2' = -10{\cdot}6183.$$

Equations (7) then give the values of $m_r'+l_r$ shown in Table 4.

Provisional values Y_{1r} and Y_{2r} are now calculated from equations (3). Thus

$$Y_{11} = -10{\cdot}3193+7{\cdot}0357 = -3{\cdot}28.$$

The working values and weights are calculated as explained above, using the appropriate table (Table XVII of Finney). It may be noted that if, as here, interpolation is required it is easiest to calculate two values of y and interpolate between these.

The remainder of the calculations proceed exactly as before, and lead to the second estimates shown in Table 4.

Table 4. *Analysis of* Drosophila *data*

Exp.	Basic data		1st estimates		2nd estimates		
	n_1	$100p_1$	y_1	W_1	Y_1	y_1	W_1
1	844	3·44	−3·37	28	−3·28	−3·349	31·2
2	2700	2·89	−3·53	81	−3·57	−3·528	75·0
3	3285	3·35	−3·37	108	−3·41	−3·378	106·8
4	1063	2·35	−3·72	26	−3·49	−3·711	31·9
5	370	8·38	−2·43	32	−2·30	−2·428	35·3
6	600	9·50	−2·30	57	−2·37	−2·301	53·6
				332			333·8
	n_2	$100p_2$	y_2	W_2	Y_2	y_2	W_2
1	1531	2·94	−3·53	46	−3·58	−3·509	42·1
2	1465	1·84	−4·01	26	−3·87	−3·979	30·3
3	4381	2·28	−3·76	97	−3·71	−3·767	106·0
4	2105	2·47	−3·68	51	−3·79	−3·683	47·0
5	550	7·82	−2·51	43	−2·60	−2·504	39·4
6	463	5·83	−2·82	27	−2·67	−2·803	31·0
				290			295·8

	l	y_1-y_2	W_1+W_2	$\frac{W_1W_2}{W_1+W_2}$	m'_r+l_r	y_1-y_2	W_1+W_2	$\frac{W_1W_2}{W_1+W_2}$	m_r+l_r	m_r
1	6·8669	+0·16	74	17·4	7·0357	+0·160	73·3	17·92	7·0543	+0·1874
2	6·8669	+0·48	107	19·7	6·6853	+0·451	105·3	21·58	6·7532	−0·1137
3	6·8669	+0·39	205	51·1	6·9062	+0·389	212·8	53·20	6·9010	+0·0341
4	6·8669	−0·04	77	17·2	6·8238	−0·028	78·9	19·00	6·8071	−0·0598
5	8·0064	+0·08	75	18·3	8·0149	+0·076	74·7	18·62	8·0134	+0·0070
6	8·0064	+0·52	84	18·3	7·9483	+0·502	84·6	19·64	7·9493	−0·0571
		+0·2990	622	142·0		+0·2937	629·6	149·96		
t_1		−10·3193				−10·3265				
t_2		−10·6183				−10·6202				

From the small changes in the values of y_1 and y_2 and the still smaller change (from +0·2990 to +0·2937) in t_1-t_2, which is trivial compared with the range of the values of which it is the weighted mean, it is clear that no further approximation is required. (A further cycle actually gives a value of +0·29341 for t_1-t_2.)

The advantage of deriving the first estimates by weighting the directly transformed proportions will now be apparent. The computations follow the same routine in each case (except for the calculation of working values) and a very satisfactory first approximation is

obtained. Moreover, the similarity of the figures in the two sets of computations provides a useful check against gross errors.*

The maximum-likelihood solution may be compared with the large-sample solution already obtained. For comparative purposes a correction to E is required to allow for the fact that squares of $\lambda\theta_1$ and $\lambda\theta_2$ have been neglected in the large-sample solution. This is obtained by multiplying each component of E by the appropriate value of $[1+\frac{1}{4}(p_1+p_2)]$. With this adjustment, and the corresponding adjustment in the weights, the large-sample estimate becomes 0·2930, with weight 147·8. The maximum-likelihood solution gives $\theta_1/\theta_2 = 1{\cdot}3414$, so that $(\theta_1-\theta_2)/\frac{1}{2}(\theta_1+\theta_2) = 0{\cdot}2916$. The agreement between the two methods of estimation is thus very close, the discrepancy being less than 1/50 of the standard error.

Estimation of error and tests of significance

Three types of test will commonly be required:

(1) A test of the difference between the two treatments, i.e. a test of whether t_1-t_2 is significantly different from zero. Whether or not the difference is significant the estimated standard error of t_1-t_2 should be given so that approximate fiducial limits can be assigned.

(2) A test whether there is any real variation in sensitivity of experiments, i.e. whether the differences between the m_r are significant.

(3) A test whether the residual variation is greater than can be accounted for by binomial sampling.

In the analysis of normally distributed data with variances in known ratios, i.e. with known relative weights, any set of parameters of which estimates have been obtained by the method of least squares can be tested for significance by finding the difference in the sum of squares accounted for when all parameters are included in the specification and that accounted for when the parameters under test are omitted from the specification. This difference is tested against the residual sum of squares by means of the z distribution. For the corresponding χ^2 test on quantal material, however, the relevant χ^2 cannot be exactly determined in this way, since changes in specification result in changes in the expectations, which in turn alter the expected variances.

A simple example will illustrate this point. Suppose we have two sets of $\frac{1}{2}n$ observations (n even), which are known to be derived from Poisson distributions whose means μ_1 and μ_2 are possibly different. If $\mu_1 = \mu_2$ the value of χ^2, χ^2_T say, based on the whole of the data will be approximately distributed as χ^2 with $n-1$ d.f. Likewise the value of χ^2, χ^2_D say, calculated from the difference of the sums of the two sets of observations, will be distributed as χ^2 with 1 d.f. Whether or not $\mu_1 = \mu_2$ variation of the observations within sets will give two values of χ^2, χ^2_1 and χ^2_2 say, each of which will be distributed as χ^2 with $\frac{1}{2}n-1$ d.f. It can easily be shown that

$$\chi^2_T = \chi^2_D + \frac{\bar{x}_1}{\bar{x}}\chi^2_1 + \frac{\bar{x}_2}{\bar{x}}\chi^2_2$$

$$= \chi^2_D + \chi^2_1 + \chi^2_2 + \frac{\bar{x}_1-\bar{x}_2}{\bar{x}}(\chi^2_1-\chi^2_2),$$

where $\bar{x}_1$ and $\bar{x}_2$ are the observed means of the two sets and $\bar{x} = \frac{1}{2}(\bar{x}_1+\bar{x}_2)$. If instead of χ^2_D the difference between the total and residual χ^2, i.e. $\chi^2_T-\chi^2_1-\chi^2_2$, is taken as the criterion

* In general we may expect to save at least a cycle of iteration by initial weighting. In Dyke and Patterson's example, for instance, the first estimates obtained in *Sampling Methods* by weighting are fully as accurate as Dyke & Patterson's second estimates.

for testing the hypothesis $\mu_1 = \mu_2$ the discrepancy given by the last term of the above equation will be obtained.

It is therefore advisable to use direct tests for each set of parameters, instead of relying on differences between χ^2 values. The general rule for obtaining the variances of the estimates of parameters in maximum-likelihood solutions of this type is to invert the matrix of the coefficients of the final set of normal equations, thus obtaining a c matrix of the customary type. The variances and covariances are then given directly by the c values.

In the present case the fact that one of the constants is redundant introduces a slight additional complication. The procedure has been given by Yates & Hale (1939). Suppose that the h normal equations $E_1 = 0$, $E_2 = 0, \ldots, E_h = 0$ are governed by the identical relationship

$$\lambda_1 E_1 + \lambda_2 E_2 + \ldots + \lambda_h E_h \equiv 0, \tag{8}$$

and that the relation

$$\mu_1 b_1 + \mu_2 b_2 + \ldots + \mu_h b_h \equiv 0 \tag{9}$$

is assumed between the estimates of the parameters. Then instead of replacing the numerical terms of the normal equations by $1, 0, \ldots, 0$ in order to obtain the equations for $c_{11}, c_{12}, \ldots, c_{1h}$, they must be replaced by

$$1 - \frac{\lambda_1\mu_1}{S(\lambda\mu)}, \quad -\frac{\lambda_1\mu_2}{S(\lambda\mu)}, \quad \ldots, \quad -\frac{\lambda_1\mu_h}{S(\lambda\mu)}.$$

Similarly the equations for $c_{21}, c_{22}, \ldots, c_{2h}$ will be obtained by replacing the numerical terms by

$$-\frac{\lambda_2\mu_1}{S(\lambda\mu)}, \quad 1 - \frac{\lambda_2\mu_2}{S(\lambda\mu)}, \quad \ldots, \quad -\frac{\lambda_2\mu_h}{S(\lambda\mu)},$$

etc.*

Designating the normal equations of our example by the suffices a, b, 1, 2, ..., k, we have

$$-E_a - E_b + E_1 + E_2 + \ldots + E_k = 0,$$

so that

$$\lambda_a = \lambda_b = -1, \quad \lambda_1 = \lambda_2 = \ldots = \lambda_k = 1.$$

Also from (4)

$$\mu_a = \mu_b = 0, \quad \mu_r = W_{1r} + W_{2r} \quad (r = 1, \ldots, k).$$

Consequently $S(\lambda\mu) = \Sigma(W_{1r} + W_{2r})$. The correction terms to be added to each of the first two sets of auxiliary equations are therefore

$$0, \quad 0, \quad g_1, \quad g_2, \quad \ldots, \quad g_k,$$

where $g_r = (W_{1r} + W_{2r})/\Sigma(W_{1r} + W_{2r})$. For the remaining k sets the same correction terms are subtracted.

Tests for the t's

The c's can now be easily evaluated. All that is necessary is to make the appropriate changes in the numerical terms of the normal equations and utilize the solutions already obtained.

* The process can be extended to two (or more) redundant constants by introducing additional correction terms $-\mu_1'\lambda_1'/S(\mu'\lambda')$, etc., provided that all the sums of products of the type $S(\lambda\mu')$, $S(\lambda'\mu)$, etc., are zero. The Yates & Hale paper is incorrect in that it omits this condition, and also assigns a numerical term μ_0 instead of zero to equation (9).

Putting $U = \Sigma \dfrac{W_{1r}W_{2r}}{W_{1r}+W_{2r}}$, we find

$$c_{aa} = V(t_1) = \frac{1}{\Sigma(W_{1r}+W_{2r})} + \frac{(\Sigma W_{2r})^2}{U\{\Sigma(W_{1r}+W_{2r})\}^2},$$

$$c_{ab} = \operatorname{cov}(t_1 t_2) = \frac{1}{\Sigma(W_{1r}+W_{2r})} - \frac{\Sigma W_{1r}\Sigma W_{2r}}{U\{\Sigma(W_{1r}+W_{2r})\}^2}.$$

Hence
$$V(t_1-t_2) = \frac{1}{U} = 1\Big/\Sigma\frac{W_{1r}W_{2r}}{W_{1r}+W_{2r}},$$

as could, indeed, be deduced directly from the form of the expression for t_1-t_2, and

$$V\tfrac{1}{2}(t_1+t_2) = \frac{1}{\Sigma(W_{1r}+W_{2r})} + \frac{\{\Sigma(W_{1r}-W_{2r})\}^2}{4U\{\Sigma W_{1r}+W_{2r}\}^2}.$$

Test for the m's

Using a similar procedure for set r, and putting

$$T_r = \frac{1}{U}\,\frac{W_{1r}\Sigma W_{2r} - W_{2r}\Sigma W_{1r}}{(W_{1r}+W_{2r})\,\Sigma(W_{1r}+W_{2r})},$$

we find
$$c_{rr} = V(m_r) = \frac{1}{W_{1r}+W_{2r}} - \frac{1}{\Sigma(W_{1r}+W_{2r})} + T_r^2,$$

$$c_{rs} = \operatorname{cov}(m_r, m_s) = -\frac{1}{\Sigma(W_{1r}+W_{2r})} + T_r T_s.$$

This gives
$$V(m_r-m_s) = \frac{1}{W_{1r}+W_{2r}} + \frac{1}{W_{1s}+W_{2s}} + (T_r-T_s)^2,$$

which, with a certain amount of algebra, can be shown to be equal to

$$1\Big/\left[\frac{W_{1r}W_{1s}}{W_{1r}+W_{1s}} + \frac{W_{2r}W_{2s}}{W_{2r}+W_{2s}}\right].$$

This provides a check, since it corresponds to the expression for $V(t_1-t_2)$ for two experiments, with interchange of treatments and experiments.

Apart from the T terms the variances and covariances of the m's are those of quantities $M_r-\bar{M}$, where the M_r are independently distributed with variances $1/(W_{1r}+W_{2r})$ and $\bar{M}$ is the weighted mean of the M_r. If the true values of the m's are all zero, therefore, the weighted sum of squares

$$\Sigma(W_{1r}+W_{2r})\,m_r^2$$

will be distributed as χ^2 with $k-1$ d.f., apart from the disturbance due to the T terms. The weighted sum of squares has the expectation

$$\Sigma(W_{1r}+W_{2r})\,V(m_r) = k-1+U\Sigma T_r^2(W_{1r}+W_{2r}).$$

If the T terms are not sufficiently small to be neglected entirely the value of the second term may be deducted before making the χ^2 test, but this correction will usually be small. Since the second term is always positive it need not in any case be calculated if significance is not attained before taking it into account.

Residual χ^2

The general formula for the residual χ^2 is the sum over all cells of

$$\frac{(\text{observed no.} - \text{expected no.})^2}{\text{expected no.}}.$$

If, as in the present case, the data consist of a set of observed proportions, this reduces to

$$S\left\{\frac{(a-m)^2}{m}+\frac{(a-m)^2}{n-m}\right\} = Sn\frac{(p-P)^2}{PQ},$$

when $a/n = p$ is the observed proportion, $m/n = P = 1-Q$ the expected proportion given by the maximum-likelihood solution, and S denotes summation over all proportions.

We may denote the provisional and working values and weights used for the sth estimates by Y_s, y_s and W_s. The provisional values Y_s will equal the expected values derived from the $(s-1)$th estimates. We have

$$y_s = Y_s + (p - P_s)\left(\frac{dY}{dP}\right)_s.$$

Consequently the residual χ^2 based on the $(s-1)$th estimates is given by

$$\begin{aligned}(\chi_r^2)_{s-1} &= Sn\frac{(y_s-Y_s)^2}{P_sQ_s}\left(\frac{dP}{dY}\right)_s^2 \\ &= SW_s(y_s-Y_s)^2. \qquad (10)\end{aligned}$$

The residual χ^2 obtained from the expected values derived from the sth estimates is given by

$$(\chi_r^2)_s = S\{W_{s+1}(y_{s+1}-Y_{s+1})^2\}. \qquad (11)$$

If the estimation process is stopped after the sth estimates it will be necessary to calculate specially Y_{s+1}, y_{s+1} and W_{s+1} to evaluate $(\chi_r^2)_s$. Usually, however, $(\chi_r^2)_{s-1}$ will be sufficiently accurate.

Finney (1952) recommends the use of the ordinary formula for the sum of the squares of the residuals when regression coefficients or constants b_1, b_2, ... are fitted by least squares. This is

$$SW(y-Y)^2 = SWy^2 - b_1B_1 - b_2B_2 - \ldots,$$

where $B_1, B_2, \ldots$ are the numerical terms of the normal equations with leading terms $b_1, b_2, \ldots$. A little consideration will show that if the sth estimates of the parameters and the numerical terms of the sth set of normal equations, together with y_s and W_s, are used to calculate the right-hand side of the above equation, we shall obtain exactly

$$SW_s(y_s-Y_{s+1})^2.$$

This, however, is not likely to give a substantially closer approximation than expression (10). The latter is easier to compute, and also has the advantage of revealing aberrant values.

When some of the expectations are very small it is well known that the χ^2 test is unreliable. In particular, positive deviations from very small expectations will make excessively large contributions to the residual χ^2. Under such circumstances Fisher (1950) has shown that $2S\{a\log_e(a/m)\}$, tested against the ordinary χ^2 distribution with the customary number of degrees of freedom, gives a much better measure of the significance of the deviations, and is well fitted to take the place of χ^2 when the expectations are small. An

example of its use is given in example (*c*) below. This criterion requires somewhat more computation for its evaluation than χ^2_r, since it cannot be simply derived from the transformed values. In our notation the contribution of each observed proportion is

$$2n\{p(\log_e p - \log_e P) + q(\log_e q - \log_e Q)\}.$$

Estimation of error and tests of significance in the *Drosophila* data

The above formulae can now be applied to the *Drosophila* data. Taking the values of the weights used for deriving the second estimates (Table 4) we have

$$V(t_1 - t_2) = 1/149{\cdot}96 = 0{\cdot}0817^2.$$

Hence $t_1 - t_2 = +0{\cdot}2937 \pm 0{\cdot}0817$. Reference to the normal probability integral gives $P = 0{\cdot}000163$. Hence the significance of the difference is clearly established. If the actual mutation rates are of interest the standard errors of t_1, t_2 and $\frac{1}{2}(t_1 + t_2)$ will also be required. These are given by

$$V(t_1) = \frac{1}{629{\cdot}6} + \frac{295{\cdot}8^2}{149{\cdot}96 \times 629{\cdot}6^2} = 0{\cdot}0553^2,$$

$$V(t_2) = \frac{1}{629{\cdot}6} + \frac{333{\cdot}8^2}{149{\cdot}96 \times 629{\cdot}6^2} = 0{\cdot}0588^2,$$

$$V\tfrac{1}{2}(t_1 + t_2) = \frac{1}{629{\cdot}6} + \frac{(333{\cdot}8 - 295{\cdot}8)^2}{4 \times 149{\cdot}96 \times 629{\cdot}6^2} = 0{\cdot}0399^2.$$

The test for variation in sensitivity between the experiments is equivalent to testing that the true values of m_r are zero. We have

$$\Sigma(W_{1r} + W_{2r})\, m_r^2 = 4{\cdot}745,$$

which has to be compared with χ^2 with 5 d.f. There is therefore no evidence of heterogeneity between experiments. Taking the T_r into account increases the expectation of the above expression by 0·0458. Clearly the T_r are of no consequence here.

Using expression (10) and the provisional and working values of the second approximation gives a value of χ^2_r of 5·15. Since the maximum-likelihood solution is very close to the minimum χ^2 solution, it is to be expected that this value of the residual χ^2 will be slightly too large, and in fact the provisional and working values of the third approximation (equivalent to the second approximation estimates) gives a value of 5·11. The value based on the provisional and working values of the second approximation is clearly adequate for practical purposes.

Other examples

As further examples, which bring out some additional points, we may take the three sets of data given by E. S. Pearson (1950).

(*a*) *Road research data*

Table 5 gives the number of cycles and motor-cycles involved in accidents causing personal injury on sections (6 to 15 miles long) of five main roads near London, together with estimates of the miles travelled by these two types of vehicle on these sections. A comparison of the probabilities of being involved in an accident per vehicle mile is required.

Since a vehicle which is involved in an accident is either removed from the road or is again exposed to risk of further accident the number of accidents will follow a Poisson distribution. If the probability of accident per vehicle mile is θ the expected number of accidents per n vehicle miles will be $n\theta$. Following the procedure of the previous example,

we may denote the probability of accident for cycles and motor-cycles on road r by $\mu_r\theta_1$ and $\mu_r\theta_2$. This will make allowances for differences between the different roads which affect both probabilities proportionally.

Table 5. *Road Research data*

Road	Cycles			Motor-cycles		
	Accidents	Miles/10^3	p	Accidents	Miles/10^3	p
1	5	1690	0·0030	4	800	0·0050
2	2	1800	0·0011	2	500	0·0040
3	3	1820	0·0016	1	330	0·0030
4	4	1790	0·0022	3	570	0·0053
5	3	1560	0·0019	5	540	0·0093
	17	8660	0·00196	15	2740	0·00547

The Poisson distribution with mean m can be regarded as the limit of the binomial distribution $(q+p)^n$, in which $p \to 0$ and $n \to \infty$ in such a manner that $np = m$. Consequently, provided the unit of distance travelled is taken sufficiently small for the chance of two accidents occurring in the same unit to be negligible, the data can be treated as if they were binomial. Here a unit of 1000 vehicle miles is sufficiently small. As before the complementary log log transformation is appropriate. Since

$$y = \log_e(-\log_e q) \to \log_e p$$

as $p \to 0$, a table of natural logarithms can be used for the actual transformation. Working values and weighting coefficients can be obtained from Finney's Table IX, which has adequate range for this purpose. Alternatively, if x is the number of occurrences a table of natural logarithms may be used with

$$W = nw = n\,e^Y,$$
$$y = Y - 1 + x/W.$$

With this procedure there is no need to choose units such that $p = x/n$ is small.

The computations follow the same lines as those of the previous example, and need not be reproduced in detail here. For cycles on the first road, for instance,

$$n = 1690, \quad n' = 5, \quad p = 0{\cdot}0030, \quad \log_e p = -5{\cdot}81,$$

and we therefore have:

	Y	y	W
1st estimate	—	−5·81	5
2nd estimate	−6·0	−5·79	4·2

the provisional value $Y = -6{\cdot}0$ being given by the first estimate of $t_1 + m_1$. For the first estimate $W = n'$, since $w = P = p$.

The values of the first and second estimates of the parameters are shown in Table 6. The second set of estimates differs little from the first and is clearly sufficiently accurate.

The average difference is clearly significant, motor-cycles being estimated as about 2·6 times as liable as cycles to be involved in an accident per 1000 vehicle miles. Taking $\pm 1{\cdot}65$ times the standard error, we obtain approximate 5% lower and upper limits of error of this ratio as 1·5 and 4·7.

The values of the various χ^2 are shown in Table 7. There is no evidence of variation in risk between the different roads, or of any variation of the difference between cycles and motor-cycles.

Table 6. *Road Research data: 1st and 2nd estimates*

	1st	2nd		1st	2nd
t_1	−6·1597	−6·1686	m_1	+0·1420	+0·1304
t_2	−5·1824	−5·1973	m_2	−0·4930	−0·4938
t_1-t_2	−0·9773	−0·9713	m_3	−0·3671	−0·3633
$\Sigma \frac{W_1 W_2}{W_1+W_2}$	7·56	7·89	m_4	−0·0020	+0·0117
S.E. (t_1-t_2)	—	±0·356	m_5	+0·2726	+0·2423

Table 7. *Road Research data: values of* χ^2

	D.F.	χ^2
Average difference (t_1-t_2)	1	7·41
Variation between roads (m's)	4	2·12
Variation in difference (residual)	4	1·40*
Total	9	10·93

* From $SW_2(Y_2-y_2)$. $SW_3(Y_3-y_3)$ gives 1·38.

That there is no appreciable variation between roads can in fact be seen by inspection of the values of p_1 and p_2 in the original data; such variation would result in a positive correlation between the values of p_1 and p_2. This fact suggests a short cut to the whole analysis, since if there is no variation between roads or in the difference between cycles and motor-cycles the data for each type of vehicle may be pooled. A combined test for absence of such variation may be quickly made from the χ^2 for the variation of the ratios for each vehicle separately. These χ^2 can be calculated in the ordinary manner and are found to be:

	D.F.	χ^2
Cycles	4	1·67
Motor-cycles	4	2·01
	8	3·68

Using the pooled data we may estimate immediately that cycles and motor-cycles have risks of 0·00196 and 0·00548 per 1000 vehicle miles of being involved in an accident. These agree closely with the estimates of 0·00209 and 0·00553 given by the previous analysis.

The possibility of pooling fragmentary data should always be examined. If pooling is justified the numerical work is considerably reduced and the presentation of the data is simplified. In addition, the amount of information is increased when heterogeneity between the different sets of data can be ignored.

(*b*) *Artificial insemination data*

These data (Table 8) are the result of an investigation (Rothschild, 1949) to test whether the passage of a small electric current for the measurement of the activity of the spermatozoa had any adverse effect on the fertility of bull semen. Seven samples were each divided into two parts, one being subjected to the electric current. The half-samples were then diluted and used to inseminate heifers. The columns headed 'returns' give the numbers of heifers which did not become pregnant, as judged by their being returned by their owners for reinsemination.

Table 8. *Artificial insemination data*

Sample	Untreated semen			Treated semen		
	Returns	Total	p	Returns	Total	p
1	5	11	0·455	1	9	0·111
2	3	9	0·333	4	9	0·444
3	3	12	0·25	3	5	0·6
4	6	11	0·545	3	9	0·333
5	4	10	0·4	0	5	0
6	3	11	0·273	1	5	0·2
7	1	4	0·25	3	6	0·5
	25	68	0·368	15	48	0·312

Inspection of the individual p values gives no indication of any heterogeneity between samples. The uniformity of the data is confirmed by the χ^2 between samples for treated and untreated semen separately. The values are:

	D.F.	χ^2
Untreated	6	3·32
Treated	6	7·92
Total	12	11·24

The data may therefore be safely pooled, giving a value of χ^2 (uncorrected) of 0·378, and corrected for continuity of 0·174, corresponding to probability levels (one tail) of 0·27 and 0·34.

The former probability may be compared with the result obtained by Rothschild, who analysed the data by means of the angular transformation, obtaining a probability (one tail) of 0·11. The poor agreement is not (as might be supposed) primarily due to the effect of pooling, but to the fact that Rothschild analysed directly transformed values without any final adjustment of the type that is automatically introduced by the use of provisional and working values. If a common provisional value is taken for the treated samples, and a second common provisional value for the untreated samples, an estimate of treated-untreated (in terms of angles) of $-4{\cdot}41 \pm 5{\cdot}50$ is obtained, giving a P (one tail) of 0·21. Similar treatment by the probit transformation also gives a P (one tail) of 0·21. This illustrates the importance of including these final adjustments not only in cases of 'special accuracy' as was implied in the earlier editions of *Statistical Tables* (Fisher & Yates), but in all cases in which their use materially affects the results.

Although there is no evidence of any harmful effect of the electric current, the treated semen being apparently the more fertile, it should be noted that the data are not sufficiently extensive to exclude all possibility of a harmful effect of some practical importance. Use of the logit transformation on the pooled data gives an upper 5% limit to treated-untreated of $+0{\cdot}23$, which is equivalent to a difference in fertility, $(1-p)$, of 0·10 in favour of the untreated semen.

(*c*) *Safety in mines data*

The data were obtained in course of an investigation at the Safety in Mines Research and Testing Branch of the Ministry of Fuel and Power, concerning the relative safety of mining explosives. Table 9 shows the result of six separate experiments comparing two explosives, A and B, under a number of different conditions. The cartridges of explosive were fixed in a gas chamber and a record was made of whether each shot resulted in an ignition or non-ignition. The weight of cartridge and the conditions of the explosion were varied in the different comparisons, but the details of these variations are not included in Pearson's paper.

Table 9. *Safety in Mines data*

Comparison	Explosive A			Explosive B		
	Ignitions	Total shots	p	Ignitions	Total shots	p
1	8	10	0·8	5	5	1·0
2	2	10	0·2	13	15	0·867
3	4	5	0·8	5	5	1·0
4	1	5	0·2	1	10	0·1
5	1	5	0·2	0	10	0
6	0	10	0	4	6	0·667

Inspection of Table 9 indicates that there is considerable heterogeneity between experiments, with Exps. 4 and 5 at one extreme and Exps. 1 and 3 at the other. The ratio of n_{1r} to n_{2r} also varies considerably from experiment to experiment. Pooling would therefore

be misleading. As in the case of the insemination data the logit transformation is most appropriate. The procedure of analysis already given in detail for the *Drosophila* data can be followed, using the table for the logit transformation. There is an initial difficulty in obtaining first estimates, however, since some values of p are 0 or 1. In these cases the transformed values are $\pm\infty$ with zero weight. These were replaced by the following (arbitrary) values; 3·0 for Exp. 5 (B), and 6 (A), in both of which 0/10 were observed, and 6·8 for 1 (B) and 3 (B), in both of which 5/5 were observed.

Table 10. *Safety in Mines data; successive estimates of* $t_1 - t_2$

Estimate	$t_1 - t_2$	Weight
1	−0·9183	9·8
2	−1·1502	10·4
3	−1·2125	8·29
4	−1·2129	7·64

With these starting values the sets of estimates shown in Table 10 were obtained. As will be seen convergence is relatively slow and three sets of estimates are required to attain satisfactory accuracy, with a fourth set to confirm that the approximation has been taken far enough. The slow convergence is not primarily due to a bad choice of starting values for the cases where $p = 0$ or 1. The main reason is the anomalous behaviour of Exps. 4 and 5. This results in very low final provisional values for explosive A in these two experiments, with correspondingly large working values and small weights. The values for Exp. 5, the more extreme of the two, in the successive approximations are as follows:

s	Y_s	y_s	W_s	Contribution to $(\chi^2_r)_s$
1	—	4·31	3·2	2·6
2	3·5	5·19	0·9	11·8
3	2·9	9·26	0·29	18·6
4	2·7	12·34	0·20	19·8
5	2·67	13·04	0·185	—

The change of y from y_3 to y_4, although apparently large, makes little difference to the estimates, largely owing to the corresponding decrease in the already small weight. The contribution to χ^2_r also reaches relative stability when $s = 3$, the change with $s = 4$ being mainly due to taking Y_4 to only one decimal place. The values of χ^2 are as follows:

	D.F.	χ^2
Average difference $(t_1 - t_2)$	1	11·3
Variation between experiments (m's)	5	33·8
Variation in difference (χ^2_r)	5	31·8
	11	76·9

The residual χ^2 is, however, unreliable in this set of data owing to the small expectations of the non-zero classes of explosive A (ignitions) in Exps. 4 and 5. The likelihood criterion $-2S\{a \log_e (a/m)\}$ has the value of 13·41, giving a significance level $P = 0{\cdot}020$. This may be compared with the significance level obtained by combining the data into the two groups, namely, Exps. 4 and 5 and Exps. 1, 2, 3 and 6. (This particular combination is valid for the purpose of a test of significance, since 4 and 5 are the two experiments with low values of n'/n.) A $2 \times 2 \times 2$ table then results. For the interaction of such a table, as Bartlett (1935) has shown, the exact probabilities of the various possible occurrences with all marginal totals fixed can be calculated in a similar manner to that followed in a 2×2 table. The resultant test is in fact a test for the constancy of the differences of the logits.* Bartlett's test gives a probability of 0·0147 of two or more of the three ignitions in Exps. 4 and 5 being associated with explosive A. Bearing in mind (*a*) that practically the whole of the discrepancy is in this test concentrated in a single degree of freedom, but that, on the other hand (*b*) there is a large effect of discontinuity, not present in the likelihood criterion, the agreement between the two tests is satisfactory.

The indication of the residual χ_r^2 that these experiments are anomalous is therefore confirmed, though the significance level is much less than that given by the χ^2 value. This result is not, however, of great consequence, since it merely provides evidence of the incorrectness of the supposition that the difference in logits between the two explosives is constant over the whole of the probability scale. Our conclusions from these data should be that explosive B produces significantly fewer ignitions than explosive A when conditions are such that the probability of ignition is substantial, but that there is no evidence of any difference when the probability is low. Substantially more data at low probabilities would be required, however, to reach any firm conclusions as to what is really happening in this region.

Summary

The analysis of sets of quantal experiments with two treatments by the use of transformations and maximum likelihood is discussed, and illustrated by examples, one of which is also treated by large-sample methods. Methods are given for (*a*) obtaining an estimate of the treatment difference in terms of the transformed variate, together with its standard error, (*b*) testing for differences in sensitivity between the different experiments, and (*c*) testing for variation in the treatment difference from experiment to experiment. The appropriateness of various transformations under different circumstances is also considered.

REFERENCES

Bartlett, M. S. (1935). *J. R. Statist. Soc. Suppl.* **2**, 248.
Bonnier, G. & Lüning, K. G. (1953). *Hereditas, Lund.*, **39**, 193.
Cochran, W. G. (1942). *Iowa State Coll. J. Sci.* **16**, 421.
Cochran, W. G. (1952). *Ann. Math. Statist.* **23**, 315.
Cochran, W. G. (1954). *Biometrics*, **10**, 417.
Dyke, G. V. & Patterson, H. D. (1952). *Biometrics*, **8**, 1.
Finney, D. J. (1952). *Statistical Method in Biological Assay*. London: Griffin.
Fisher, R. A. (1932). *Statistical Methods for Research Workers* (4th ed.). Edinburgh: Oliver and Boyd.

* Simpson (1951), following a suggestion by Bartlett, has drawn attention to certain errors in the latter's paper, which do not, however, affect the test for interaction. Lancaster (1951) has criticized Bartlett's test on what seems to me the invalid ground that the components of χ^2 associated with this test are not additive.

FISHER, R. A. (1950). *Biometrics*, **6**, 17.
FISHER, R. A. & YATES, F. (1938). *Statistical Tables for Biological, Agricultural and Medical Research* (4th ed. 1953). Edinburgh: Oliver and Boyd.
JOLLY, G. M. (1950). *Ann. Appl. Biol.* **37**, 597.
LANCASTER, H. O. (1949). *Biometrika*, **36**, 370.
LANCASTER, H. O. (1951). *J. R. Statist. Soc.* B, **13**, 242.
PEARSON, E. S. (1950). *Biometrika*, **37**, 383.
PEARSON, K. (1933). *Biometrika*, **25**, 379.
ROTHSCHILD, LORD (1949). *J. Agric. Sci.* **39**, 294.
SIMPSON, E. H. (1951). *J. R. Statist. Soc.* B, **13**, 238.
YATES, F. (1953). *Sampling Methods for Censuses and Surveys* (2nd ed.). London: Griffin.
YATES, F. (1955). *Biometrika*, **42**, 404.
YATES, F. & HALE, R. W. (1939). *J. R. Statist. Soc. Suppl.* **6**, 67.

A NOTE ON THE APPLICATION OF THE COMBINATION OF PROBABILITIES TEST TO A SET OF 2×2 TABLES

By F. YATES, F.R.S.

Rothamsted Experimental Station

Introduction

The combination of probabilities test, or probability integral test, as it is also called, has been commonly used to give an overall test of significance for a set of 2×2 contingency tables for which pooling is not regarded as admissible because of heterogeneity of the data. Data of this type can, in my opinion, be more effectively analysed by the method of maximum likelihood, using some appropriate transformation; the methods of analysis are described in a separate paper (Yates, 1955). Cases will, however, arise where a quick, possibly preliminary, test of significance is required, and for this purpose the combination of probabilities test or some analogous test may be regarded as adequate. It therefore seems worth putting on record certain points concerning tests of this type.

The combination of probabilities test requires the calculation of a significance level P for each of the k tables separately. The test criterion $S(-2\log_e P)$ is then compared with the χ^2 distribution with $2k$ degrees of freedom. We usually require to test whether the data collectively show evidence of deviation in a common direction, and in such cases the significance level P must be that of the appropriate single tail of the corresponding distribution.

The test is exact if the basic P distributions are continuous with uniform probability, i.e. if each P can be regarded as uniformly distributed between 0 and 1 when the hypothesis to be tested holds. If the P's are derived from 2×2 contingency tables with small marginal totals, however, the distributions of P are by no means continuous. As has been pointed out by Pearson (1950) they can be made continuous with uniform probability by the following device. If for any particular 2×2 table the probability derived from the hypergeometric series is $P_a(H)$ that the number in a particular cell is $\geqslant a$, and the probability of this number being $\geqslant (a+1)$ is $P_{a+1}(H)$, then if a is the number observed we may use a value $P_{a+\theta}(H)$ instead of $P_a(H)$, where $P_{a+\theta}(H)$ is a random selection with uniform probability from values in the range $P_a(H)$ to $P_{a+1}(H)$.

Since all the values of P will be decreased by applying this adjustment, the value of $S(-2\log_e P)$ will be increased. Without the adjustment, therefore, we shall consistently over-estimate the combined probability of obtaining a set of P's whose product is as small as or smaller than that observed. This confirms the result obtained by Lancaster (1949) in the course of a detailed investigation on the combination of probabilities from discrete distributions.

To overcome this difficulty Lancaster suggested that the average value of $-2\log_e P_{a+\theta}(H)$ be used instead of the value $-2\log_e P_a(H)$. This average value he denoted by χ^2_m, but to avoid confusion with the various χ^2 with 1 d.f. appertaining to the individual experiments, we will here denote it by $-2\log_e P_{a+}(H)$. We have

$$\begin{aligned} -2\log_e P_{a+}(H) = \chi^2_m &= E\{-2\log_e P_{a+\theta}(H)\} \\ &= 2 - 2\{P_a(H)\log_e P_a(H) - P_{a+1}(H)\log_e P_{a+1}(H)\}/\{P_a(H) - P_{a+1}(H)\}. \end{aligned}$$

As a further alternative, to save computation, Lancaster suggested what he called the median value χ^2, which he defined by

$$\chi'^2_m = -2\log_e \tfrac{1}{2}\{P_a(H)+P_{a+1}(H)\} \quad P_{a+1}(H)\neq 0,$$
$$= 2-2\log_e P_a(H), \quad P_{a+1}(H)=0.$$

Neither of Lancaster's procedures, of course, gives an exact χ^2 distribution with $2k$ d.f., but the departures will be very slight in all cases likely to be met with in practice. By Pearson's procedure an exact χ^2 distribution can be generated, but this introduces an arbitrary element which, as Pearson himself recognized, may be regarded as altogether too high a price to pay for a formally exact test.

USE OF χ^2 TO DETERMINE THE VALUES OF P

If the number of terms in the hypergeometric series appertaining to a single experiment is at all large the calculation of the exact probabilities is a tedious matter. An alternative procedure is to calculate the values of χ^2 for the individual experiments without the correction for continuity, and use these to determine the corresponding one-tail probabilities. Cochran (1942, 1952), correcting a misleading statement by myself (1934), pointed out that the continuity correction should not be applied when summing χ^2 from a number of 2×2 tables, though the reason he gives in his 1952 paper, that 'the correction has a tendency to over-correct' and that 'the over-correction mounts up in a disconcerting manner' is not, as we have seen above, the basic one. Lancaster discussed the use of χ^2 without correction for continuity for the combination of probabilities test, but confined his attention to the calculation of the two-tail probabilities. Unfortunately, as already mentioned, this is not what is usually required in practice, and Lancaster's argument is not applicable to the one-tail case.

The question of the accuracy of the χ^2 approximation, therefore, still requires investigation. When χ^2 corrected for continuity, χ^2_c say, gives a good approximation to the one-tail probabilities, it may be expected that the use of χ^2 without correction for continuity, χ^2_u say, for the calculation of the probabilities for combination will give a reasonable approximation to the test based on the exact probabilities and $-2\log_e P_{a+}(H)$ as defined above. For if $P_a(\chi_u)$ is the one-tail probability given by χ^2_u, when the value a is observed, and $P_a(\chi_c)$ is the corresponding probability given by χ^2_c, then $P_a(\chi_u)$ is intermediate between $P_a(\chi_c)$ and $P_{a+1}(\chi_c)$, and will not differ greatly from $P_{a+}(\chi_c)$.

It is well known, however, that if the relevant hypergeometric distribution is very asymmetrical $P_a(\chi_c)$ is not a good approximation to $P_a(H)$ (Yates, 1934). Consequently, if the data are such that the hypergeometric distributions for the separate tables are very asymmetrical, and particularly if the asymmetry is in the same sense for all tables, the use of χ^2 without correction for continuity instead of the exact probabilities may be expected to give misleading results.

Some comparisons made for actual 2×2 tables, however, are reassuring. Tables having the following marginal totals (treatments 1 and 2) were taken:

	Table A			Table B			Table C		
	1	2		1	2		1	2	
+	a		8	a		8	a		16
−			192			142			224
	100	100	200	50	100	150	40	200	240

Table A gives a symmetrical distribution and Tables B and C give increasing degrees of asymmetry. (No particular merit is claimed for these tables; A was originally chosen for another purpose, and B and C were taken to give distributions differing from A mainly in degree of symmetry.)

For these three tables the values of $-2\log_e P_{a+}(H)$ and $-2\log_e P_a(\chi_u)$ were calculated for all possible values of a. For Tables B and C there are two sets of values depending on the direction in which the deviations are measured. The results are shown in Table 1. In distribution A all the individual values except that for $a = 8$ (which will only occur in 1 out

Table 1. *Comparison of* $-2\log_e P_{a+}(H)$ *and* $-2\log_e P_a(\chi_u)$

Distribution A

a	$p(a)$	Test for 1 > 2	
		$-2\log_e P_{a+}(H)$	$-2\log_e P_a(\chi_u)$
0	0·003377	0·004	0·004
1	0·029052	0·036	0·030
2	0·107092	0·181	0·155
3	0·220948	0·583	0·536
4	0·279062	1·413	1·386
5	0·220948	2·842	2·894
6	0·107092	5·055	5·195
7	0·029052	8·332	8·382
8	0·003377	13·383	12·484
	1·000000	2·000	2·004

Distribution B

a	$p(a)$	Test for 1 > 2		Test for 1 < 2	
		$-2\log_e P_{a+}(H)$	$-2\log_e P_a(\chi_u)$	$-2\log_e P_{(8-a)+}(H)$	$-2\log_e P_{(8-a)}(\chi_u)$
0	0·035397	0·036	0·040	8·682	7·833
1	0·152244	0·239	0·210	4·571	4·616
2	0·277764	0·805	0·724	2·302	2·384
3	0·280688	1·905	1·839	1·021	1·017
4	0·171775	3·662	3·767	0·371	0·330
5	0·065168	6·178	6·646	0·102	0·073
6	0·014962	9·581	10·56	0·019	0·010
7	0·001900	14·11	15·56	0·002	0·001
8	0·000102	20·38	21·67	0·000	0·000
	1·000000	2·000	2·021	2·000	1·989

Table 1. (*cont.*)

Distribution C

a	$p(a)$	Test for 1>2		Test for 1<2	
		$-2\log_e P_{a+}(H)$	$-2\log_e P_a(\chi_u)$	$-2\log_e P_{(18-a)+}(H)$	$-2\log_e P_{(18-a)}(\chi_u)$
0	0·048696	0·049	0·065	8·044	6·881
1	0·168463	0·288	0·264	4·190	4·182
2	0·264922	0·874	0·776	2·152	2·268
3	0·251227	1·907	1·791	1·011	1·050
4	0·160692	3·426	3·460	0·416	0·390
5	0·073459	5·442	5·890	0·144	0·108
6	0·024809	7·955	9·148	0·041	0·021
7	0·006309	10·96	13·27	0·009	0·002
8	0·001220	14·46	18·29	0·002	0·000
9	0·000180	18·47	24·23	...	...
10	0·000020	...	...	...	...
11	0·000002	...	...	...	...
...	...	...	...	...	...
	0·999999	2·000	2·030	2·000	1·975

of 300 experiments) agree to within 0·15, and the expectation of $-2\log_e P_a(\chi_u)$ is almost exactly 2. Distributions B and C show considerably greater individual differences, as is to be expected, but the large differences are again confined to the tails, where they are of no great consequence; for if values occur in the tail in any substantial proportion of the experiments the significance of the combined results will not in any case be in doubt. The expectations deviate in opposite directions for the two types of test but are still very close to 2.

On this last point it may be noted that Lancaster defines (without comment) what he calls the crude χ^2 of a 2×2 table as

$$\chi^2 = \frac{(N-1)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

The customary form, which has been used in this paper for χ_u^2, has the factor N instead of $N-1$ in the numerator. Lancaster's form has the property, which is useful for the combination of two-tail tests, that the expectation is 1. This, of course, does not make the expectation of $-2\log_e P_a(\chi)$ equal to 2, but it does in fact bring it closer to 2 in the case of symmetrical distributions. It can do little, however, to improve the expectations for asymmetrical distributions, since the deviations are of opposite sign for the two types of test.

From the above comparisons we may conclude that the determination of the one-tail probabilities from χ^2 without correction for continuity will be satisfactory in cases likely to be met with in practice, even when the expectations in the individual experiments are quite small.

Variants of the test

The use of values of χ^2 for 2 d.f. for the combination of probabilities is to a certain extent arbitrary. It has the convenience that the values are easily calculated, and the use of a function of the product of the probabilities has a certain intuitive appeal, but the method would work equally well with other basic numbers of degrees of freedom. If, for instance, the values of χ^2 for 1 d.f. corresponding to the P's are summed then in the absence of association the sum will be distributed as χ^2 for k d.f. Equally—and in the type of data we are considering this procedure requires even less computation—the values of χ_u in the absence of association approximate to normal deviates with zero mean and unit standard deviation, and their sum is therefore a normal deviate with a standard deviation of $\sqrt{k}$ (see, for example, Cochran, 1954).

The calculations for these variants for the *Drosophila* data analysed in the paper referred to (Yates, 1955) are exhibited in Table 2. The combination of probabilities test (column 4) gives a significance level (one tail) of $P = 0{\cdot}000610$, whereas the use of χ^2 for 1 d.f. instead of 2 d.f. (column 5) gives $P = 0{\cdot}000760$, and $S(\chi_u)$ (column 2) gives $P = 0{\cdot}000673$.

Table 2. *Calculation of significance levels for the* Drosophila *data*

Exp.	χ^2_u	χ_u	$P_a(\chi_u)$	$-2 \log_e P_a(\chi_u)$	$\chi^2_{(1)}\{P_a(\chi_u)\}$
	(1)	(2)	(3)	(4)	(5)
1	0·4448	0·667	0·2524	2·753	1·309
2	4·2274	2·056	0·01989	7·835	5·420
3	8·0065	2·830	0·002327	12·126	9·272
4	0·0418	−0·204	0·5808	1·087	0·305
5	0·0938	0·306	0·3798	1·936	0·771
6	4·8325	2·198	0·01397	8·542	6·042
	17·6468	7·853	—	34·279	23·119
P	0·00718	0·000673	—	0·000610	0·000760

These values are in close agreement. Indeed, except when the data from the different experiments are mutually contradictory no great differences are to be expected between the different possible methods that suggest themselves, since they merely demarcate equiprobable contour surfaces in the multiple P space which are not markedly different. Nevertheless, alternative tests which are merely variants based on the same information are likely to give rise to confusion, and it is therefore recommended that some standard procedure is adopted whenever a test based on the levels of significance in the individual experiments or sets of observations is required. If the information presented to the statistician consists of the relevant probabilities the combination of probabilities test commends itself both on historical grounds and because of its simplicity and intuitive appeal. If a one-tail test of a set of 2 × 2 tables based on the χ^2 values for each table is required, however, there seems little point in transforming the values of χ_u into probabilities and then retransforming these into χ^2 values. It is therefore recommended that in such cases $S(\chi_u)$ be adopted as the standard test criterion.

In contrast with the above values of P it may be noted that $S(\chi_u^2)$ (column 1), which in the absence of association would also be approximately distributed as χ^2 with 6 d.f., gives a value of $P = 0{\cdot}00718$. This considerably larger value of P is to be expected, since the direction of the deviations is not taken into account.

EFFICIENCY OF THE TEST

The significance levels given by the combination of probabilities test and by the ratio of $t_1 - t_2$ to its standard error in the maximum-likelihood solution for the four examples of the other paper (Yates, 1955) are as follows:

	Combination of probabilities	Maximum likelihood
Drosophila data	0·000610	0·000163
Road research data	0·00691*	0·00317
Artificial insemination data	0·64*	0·79
Safety in mines data	0·000517*	0·000402

* Computed from the values of $S(\chi_m^2)$ given by Pearson (1950).

In all three cases in which significance is attained the maximum-likelihood solution gives a higher level of significance than the combination of probabilities test. In two cases the difference is substantial. This suggests that the combination of probabilities test is not very efficient, though the differences may in part be due to chance causes or to inadmissible approximations in one or both the tests.

Lancaster compared the power of the combination of probabilities test, the direct summation of χ^2 test, and the test derived from the pooled data by numerical methods, taking the case of the binomial distribution $(q+p)^5$ with $p = 0{\cdot}5$ as the null hypothesis and enumerating all possible events. He made a similar comparison of the first two tests and some variants on a 2×2 table. Unfortunately, however, he used two-sided tests. This considerably reduces the efficiency of the combination of probabilities or summation of χ^2 tests, and his estimates of the relative power of the three tests are therefore not really relevant to the situations met with in practice.

A full investigation of the efficiency of the combination of probabilities test is beyond the scope of this note, but it is to be expected on general grounds that the combination of probabilities test will be somewhat inefficient. In the first place no direct cognizance is taken of the number of observations and other factors that affect the amount of information given by the different experiments. It is true that the more accurate experiments will, on the average, yield results of greater significance when there is a real difference. But if, for example, of two experiments A and B, A yields 10 times the amount of information given by B, we should naturally be inclined to give more weight to a significant result from experiment A than to one from experiment B. But on the combination of probabilities test the result $P_A = 0{\cdot}025$, $P_B = 0{\cdot}3$, for instance, will give exactly the same combined level of significance (5 % single tail), as $P_A = 0{\cdot}3$, $P_B = 0{\cdot}025$.

In the second place even if all the experiments yield the same amount of information this information is not in general combined efficiently by the combination of probabilities test.

As a simple example we may consider the case of a set of quantitative experiments which furnish normally distributed estimates $x_1, x_2, \ldots, x_k$ of a constant treatment difference μ, all having the same (known) variance σ^2. The efficient (indeed sufficient) combined estimate of μ will then be $\bar{x}$, and this will be normally distributed with known variance σ^2/k and can therefore be used to give an exact overall P, P_L say. The exact level of significance can also be calculated for each experiment separately, and these levels can then be combined to give an overall P, P_C say. It is easily seen that in particular cases P_C may differ considerably from P_L.

Table 3. *Comparison of significance levels given by maximum likelihood (P_L) and combination of probabilities (P_C) in a sampling experiment*

$P(\bar{x}-0{\cdot}9)$	P_L	P_C	$\log_{10}(P_C/P_L)$
0·9987	0·839	0·820	−0·010
0·729	0·0808	0·0955	0·073
0·719	0·0764	0·0460	−0·220
0·702	0·0694	0·0908	0·117
0·663	0·0559	0·0409	−0·136
0·659	0·0548	0·0296	−0·267
0·641	0·0495	0·0512	0·015
0·618	0·0436	0·0838	0·284
0·540	0·0281	0·0145	−0·287
0·480	0·0197	0·0367	0·270
0·405	0·0122	0·0318	0·416
0·405	0·0122	0·0132	0·034
0·334	0·00734	0·0216	0·469
0·258	0·00391	0·0100	0·408
0·233	0·00307	0·00394	0·108
0·142	0·00104	0·00118	0·055
0·131	0·000874	0·00185	0·326
0·111	0·000619	0·00287	0·666
0·090	0·000404	0·00132	0·514
0·038	0·0000784	0·0000973	0·094
			0·146

A test of the relative performance of the two tests was made for a specific example of this type with $k = 5$, $\sigma = 1$, and a true treatment difference of 0·9. With these values $\bar{x}$ will attain significance at the 2·2 % level (single tail) in 50 % of all sets of five experiments. The actual levels attained by $\bar{x}$ and by the combination of probabilities test were calculated for twenty such sets, using random normal deviates nos. 6551–6650, each increased by 0·9, given by Wold (1948). The results are shown in Table 3. $\bar{x} - 0{\cdot}9$ will be normally distributed about zero mean with standard deviation $1/\sqrt{5}$. The probability $P(\bar{x} - 0{\cdot}9)$ of getting a value of $\bar{x} - 0{\cdot}9$ greater than the observed value was calculated for each set, and the results have been arranged in order of magnitude of this probability. With a large number of sets there would be a uniform distribution of $P(\bar{x} - 0{\cdot}9)$ over the range 0–1.

The results show clearly that P_L tends to be less than P_C, and that the difference is greater in the experiments with small $P(\bar{x} - 0{\cdot}9)$ which, of course, also have small P_L. The average

value of $\log_{10} P_C/P_L$ is 0·146. In other words, the significance levels are on the average in the ratio of 1·40 : 1. There is thus a clear loss of efficiency with P_C. Equally important, there are considerable discrepancies in individual sets of experiments, leading to a conflict of evidence from the same body of data which arises solely from the use of an inefficient test.

SUMMARY

It is shown that χ^2 without correction for continuity will give one-tail probabilities for 2×2 tables which may be safely combined in most cases likely to be met with in practice. The summation of the corresponding signed values of χ gives a rapid method of combination. Reasons are given for believing that combination of probabilities tests are not likely to be very efficient, and this conclusion is demonstrated by a small sampling experiment.

REFERENCES

COCHRAN, W. G. (1942). *Iowa State Coll. J. Sci.* **16**, 421.
COCHRAN, W. G. (1952). *Ann. Math. Statist.* **23**, 315.
COCHRAN, W. G. (1954). *Biometrics*, **10**, 417.
LANCASTER, H. O. (1949). *Biometrika*, **36**, 370.
PEARSON, E. S. (1950). *Biometrika*, **37**, 383.
WOLD, H. (1948). *Random Normal Deviates*. Tracts for computers, no. XXV. Cambridge University Press.
YATES, F. (1934). *J. R. Statist. Soc. Suppl.* **1**, 217.
YATES, F. (1955). *Biometrika*, **42**, 382.

A TEST FOR HOMOGENEITY OF THE MARGINAL DISTRIBUTIONS IN A TWO-WAY CLASSIFICATION

By ALAN STUART

Division of Research Techniques, London School of Economics

1. Introduction

There are several circumstances in which we may wish to test the homogeneity of the two sets of marginal probabilities in a two-way classification. For example, a sample from a bivariate distribution (say height of father, height of son) may be classified into a two-way table with identical (height) groupings in each margin. Or a similar classification may be possible for a non-measurable variable (say strength of right hand – strength of left hand). Again, in surveys of the same sample (a 'panel') on two different occasions, the interrelation of the results on the two occasions may be displayed in a two-way table, with one margin corresponding to each occasion. In all these cases, the question may arise: are the two sets of marginal probabilities identical?

If the variable is measurable, we may test the difference between the *means* of the two marginal distributions by a large-sample standard-error test. However, we may be interested in the overall distributions, rather than only in their means. For the more stringent hypothesis of homogeneity, a test exists if we have two completely independent samples, when an ordinary χ^2 test of homogeneity may be applied (Cramer, 1946, p. 445). This test does not meet the essentially bivariate situations described above, where non-independence of the marginal distributions is a fundamental feature of the problem.

When the classification is a double dichotomy, the problem of testing marginal homogeneity is simple, and its solution is a special case of the large-sample solution of the more general 2^K classification problem given by Cochran (1950). Bowker (1948) gave a large-sample test for complete symmetry in a two-way classification, a more restrictive hypothesis which is concerned with the entire set of probabilities in the classification, and not only with the marginal probabilities as we are here. In the present paper, a large-sample test for marginal homogeneity is derived and illustrated.

2. Variances and covariances

With the usual notation for an $m \times m$ table, we denote the probability of falling into the ith row and jth column by p_{ij}, and define the marginal probabilities by

$$p_{i.} = \sum_j p_{ij}, \quad p_{.j} = \sum_i p_{ij},$$

while obviously

$$\sum_i \sum_j p_{ij} = \sum_i p_{i.} = \sum_j p_{.j} = 1.$$

The corresponding sample numbers are denoted by n_{ij}, $n_{i.}$, $n_{.j}$, while the total sample size is n.

By standard multinomial theory, we have for the means, variances and co-variances of the n_{ij}

$$\left.\begin{aligned} E(n_{ij}) &= np_{ij}, \\ V(n_{ij}) &= np_{ij}(1-p_{ij}), \\ C(n_{ik}, n_{lj}) &= -np_{ik}p_{lj} \quad (i \neq l \text{ and/or } k \neq j), \end{aligned}\right\} \tag{1}$$

and also

$$\left.\begin{aligned} E(n_{i.}) &= np_{i.}, \\ V(n_{i.}) &= np_{i.}(1-p_{i.}), \\ C(n_{i.}, n_{j.}) &= -np_{i.}p_{j.} \quad (i \neq j), \\ C(n_{.i}, n_{.j}) &= -np_{.i}p_{.j} \quad (i \neq j). \end{aligned}\right\} \tag{2}$$

We now require

$$C(n_{i.}, n_{.j}) = V(n_{ij}) + \sum_{k,l} C(n_{ik}, n_{lj}),$$

where $i \neq l$ and/or $k \neq j$. From (1), this is

$$\begin{aligned} C(n_{i.}, n_{.j}) &= n\{p_{ij}(1-p_{ij}) - \sum_{k,l} p_{ik}p_{lj}\}, \\ &= n(p_{ij} - p_{i.}p_{.j}), \end{aligned} \tag{3}$$

on taking the term in p_{ij}^2 into the summation. We now define the statistics

$$d_i = n_{i.} - n_{.i} \quad (i = 1, 2, \ldots, m).$$

Since the likelihood-ratio principle yields an intractable result in this case, it seems natural to use the d_i as the basis for a test concerning the differences between the corresponding marginal probabilities. For their means, variances and co-variances, we have, from (2) and (3), the exact results:

$$E(d_i) = E(n_{i.} - n_{.i}) = n(p_{i.} - p_{.i}), \tag{4}$$

$$\begin{aligned} V(d_i) &= V(n_{i.}) + V(n_{.i}) - 2C(n_{i.}, n_{.i}) \\ &= n\{(p_{i.} + p_{.i} - 2p_{ii}) - (p_{i.} - p_{.i})^2\}. \end{aligned} \tag{5}$$

Also, for $i \neq j$

$$\begin{aligned} C(d_i, d_j) &= C(n_{i.}, n_{j.}) + C(n_{.i}, n_{.j}) - \{C(n_{i.}, n_{.j}) + C(n_{.i}, n_{j.})\} \\ &= -n\{(p_{ij} + p_{ji}) + (p_{i.} - p_{.i})(p_{j.} - p_{.j})\}. \end{aligned} \tag{6}$$

3. TESTING THE HYPOTHESIS OF HOMOGENEITY

Suppose that we wish to test the hypothesis

$$H_1: \quad (p_{i.} - p_{.i}) = \Delta_i \quad (i = 1 \text{ to } m),$$

where, of course, we must have

$$\sum_{i=1}^{n} \Delta_i = 0.$$

(4), (5) and (6) then become

$$\left.\begin{aligned} E(d_i \mid H_1) &= n\Delta_i, \\ V(d_i \mid H_1) &= n\{(p_{i.} + p_{.i} - 2p_{ii}) - \Delta_i^2\}, \\ C(d_i, d_j \mid H_1) &= -n\{(p_{ij} + p_{ji}) + \Delta_i\Delta_j\} \quad (i \neq j), \end{aligned}\right\} \tag{7}$$

and in the special case of particular interest here, where we have

$$H_0: \quad \Delta_i = 0 \quad (\text{all } i),$$

(7) becomes

$$\left.\begin{aligned} E(d_i \mid H_0) &= 0, \\ V_{ii} = V(d_i \mid H_0) &= n(p_{i.} + p_{.i} - 2p_{ii}), \\ V_{ij} = C(d_i, d_j \mid H_0) &= -n(p_{ij} + p_{ji}) \quad (i \neq j). \end{aligned}\right\} \tag{8}$$

We have deliberately refrained from replacing $p_{.i}$ by $p_{i.}$ in V_{ii}, as we could have done, since if, as will generally be the case, the true probabilities are unknown, the maximum-likelihood estimators of the variances and co-variances in (8) are

$$\left.\begin{aligned} \hat{V}_{ii} &= n_{i.} + n_{.i} - 2n_{ii}, \\ \hat{V}_{ij} &= -(n_{ij} + n_{ji}). \end{aligned}\right\} \tag{9}$$

It is well known that the n_{ij} have a limiting multinormal distribution with moments given by (1). It follows that the $n_{i.}$ and $n_{.i}$ will also be jointly normally distributed, since they are linear functions of the n_{ij}, and finally, by the same argument, that the variates d_i ($i = 1$ to m) will be asymptotically multinormally distributed with moments given by (7) in general, and by (8) on the null hypothesis. To a further degree of approximation, the same result holds if (9) is used as an estimator of (8).

Now it is a standard result that in the exponent, say $-\frac{1}{2}Q$, of a multinormal distribution, Q is distributed like χ^2 with degrees of freedom equal in number to the rank of the distribution. Furthermore, Q in this exponent is simply a quadratic form in the variables.

Thus, on H_0,

$$Q = \sum_{i,j=1}^{m} a_{ij} d_i d_j \tag{10}$$

is distributed in the limit like χ^2 with $(m-1)$ degrees of freedom, the rank of the distribution being $(m-1)$ since $\sum_{i=1}^{n} d_i = 0$, and in general this is the only constraint on the d_i. If the distribution were non-singular, the coefficients a_{ij} would simply be the elements of the inverse of the variance-covariance matrix (V_{ij}). As a result of the singularity, (V_{ij}) itself is also singular, and so we cannot invert it.

However, any marginal distribution of a multinormal distribution is itself normal, so that we may eliminate the redundant variate (say the last) and obtain the result that

$$Q = \sum_{i,j=1}^{m-1} V^{ij} d_i d_j \tag{11}$$

is distributed in the limit like χ^2 with $(m-1)$ degrees of freedom, $(m-1)$ still being the rank of the distribution. If we replace the V^{ij} in (11) by their maximum-likelihood estimators given by (9), the same asymptotic result holds. In (11), (V_{ij}) is now understood to be defined for $i, j = 1$, to $m-1$ only, and (V^{ij}) is its inverse.

The fact that (11) takes no explicit account of d_m lends it an appearance of arbitrariness, but although the values of the terms in Q are changed by eliminating some other d_i instead of d_m, their sum Q is uniquely determined. This is in virtue of the fact that since any $(m-1)$ of the d_i uniquely determine the other one, Q must be, apart from constants, the log likelihood of the complete set of d_i, irrespective of which of the d_i is omitted in Q. In point of fact, Q could be expressed as a function of all the m values of d_i and the complete matrix

(V_{ij}), but this merely complicates the computational procedure and makes no difference to the result.

Finally, we must determine the appropriate critical region of the distribution of Q. (7) shows that, when H_0 is not true, the variance-covariance matrix of the d_i may be written

$$(V_{ij}) = n(w_{ij}),$$

where (w_{ij}) is independent of n. The inverse matrix is therefore

$$(V^{ij}) = \frac{1}{n}(w^{ij}).$$

Thus the expected value of Q in (11) is

$$E(Q \mid H_1) = \sum_{i,j=1}^{m-1} \frac{1}{n} w^{ij} E(d_i d_j)$$

$$= \sum \frac{1}{n} w^{ij}\{C(d_i, d_j) + E(d_i)\,E(d_j)\},$$

so that, using (7),

$$E(Q \mid H_1) = \mathrm{O}(n). \tag{12}$$

On the other hand, the limiting χ^2 distribution of Q on H_0 gives us

$$E(Q \mid H_0) = m - 1, \tag{13}$$

independent of n. Thus, with increasing n, the difference between (12) and (13) will exceed any bound, and the appropriate critical region for our large-sample test is the upper tail of the distribution of Q. The same conclusion holds if (V_{ij}) is replaced by its maximum-likelihood estimator $(\hat{V}_{ij})$.

4. COMPUTATIONAL PROCEDURE: AN EXAMPLE

We now set out the computation necessary for the test, and give an illustration of its use.

(1) Form the matrix $(\hat{V}_{ij})$ given by (9), for $i, j = 1$ to $(m-1)$.

(2) Invert $(\hat{V}_{ij})$ to obtain $(\hat{V}^{ij})$.

(3) Compute $Q = \sum_{i,j=1}^{m-1} \hat{V}^{ij} d_i d_j$, where $d_i = n_{i.} - n_{.i}$, and test in the χ^2 distribution with $(m-1)$ degrees of freedom, the critical region being the upper tail.

Consider the following example, which uses data quoted by Stuart (1953).

7477 women aged 30–39; unaided distance vision

Left eye / Right eye	Highest grade	Second grade	Third grade	Lowest grade	Total
Highest grade	1520	266	124	66	1976
Second grade	234	1512	432	78	2256
Third grade	117	362	1772	205	2456
Lowest grade	36	82	179	492	789
Total	1907	2222	2507	841	7477

We have

$$d_1 = 1976 - 1907 = 69,$$
$$d_2 = 2256 - 2222 = 34,$$
$$d_3 = 2456 - 2507 = -51,$$

and we check that $d_4 = 789 - 841 = -52$ equals the sum $(d_1 + d_2 + d_3)$, negatively signed. Since d_1 and d_2 are positive, while d_3 and d_4 are negative, one may ask whether the sight of the right eye may really be regarded as better than that of the left eye for the population from which this is a sample. The estimated variance matrix of d_1, d_2, d_3 is obtained by use of (9), as

$$\hat{V}_{11} = (266 + 124 + 66) + (234 + 117 + 36) = 843,$$
$$\hat{V}_{12} = \hat{V}_{21} = -(234 + 266) = -500,$$

and so on, giving

$$(\hat{V}_{ij}) = \begin{pmatrix} 843 & -500 & -241 \\ -500 & 1454 & -794 \\ -241 & -794 & 1419 \end{pmatrix}.$$

The inverse is obtained directly as

$$(\hat{V}^{ij}) = 10^{-6}\begin{pmatrix} 2482 & 1560 & 1295 \\ 1560 & 1972 & 1368 \\ 1295 & 1368 & 1690 \end{pmatrix},$$

and we have, for our χ^2 statistic with 3 degrees of freedom,

$$Q = 10^{-6}\{(2482 \times 69^2) + (1972 \times 34^2) + (1690 \times 51^2) + 2(1560 \times 69 \times 34) - 2(1295 \times 69 \times 51) - 2(1368 \times 34 \times 51)\}$$
$$= 11{\cdot}96.$$

The 1 % point for the χ^2 distribution with 3 degrees of freedom is 11·345, so we conclude that the result is significant of a difference in the distribution of sight in right and left eye.

If, instead, of eliminating d_4, we had eliminated, say, d_1 we should have obtained as above:

$$(\hat{V}_{ij}) = \begin{pmatrix} 1454 & -794 & -160 \\ -794 & 1419 & -384 \\ -160 & -384 & 646 \end{pmatrix},$$

$$(\hat{V}^{ij}) = 10^{-6}\begin{pmatrix} 1332 & 995 & 921 \\ 995 & 1583 & 1187 \\ 921 & 1187 & 2482 \end{pmatrix},$$

and finally

$$Q = 10^{-6}\{(1332 \times 34^2) + (1583 \times 51^2) + (2482 \times 52^2) + 2(1187 \times 51 \times 52) - 2(995 \times 34 \times 51) - 2(921 \times 34 \times 52)\}$$
$$= 11{\cdot}96$$

as before.

I am grateful to Mr J. Durbin for several illuminating discussions of this problem.

REFERENCES

Bowker, Albert H. (1948). A test for symmetry in contingency tables. *J. Amer. Statist. Ass.* **43**, 572–4.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, **37**, 256–66.

Cramer, Harald (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika*, **40**, 105–10.

DISTRIBUTIONS OF KENDALL'S TAU BASED ON PARTIALLY ORDERED SYSTEMS

BY SOL HABERMAN

University of Minnesota

INTRODUCTION

If a pair of objects a, b are connected by a relation R which is non-reflexive, anti-symmetric and transitive, we say, if aRb, that *a precedes b*. Suppose we have a set of n objects in which R holds between some, not necessarily all, objects. The situation may be represented diagrammatically by a set of points, one to each object, joined by lines between each pair of objects which are connected by R. The relation aRb between any pair exists, if any, only if there is a path along the network joining the corresponding points. We may represent the direction of the relationship by an arrow on the corresponding line, and the relation aRb requires that the arrows along one path, at least between a and b, all go in the same direction. For convenience, in this paper, we shall regard all our directions as going downwards (from the head of the page towards the foot), and it will not be necessary actually to draw the arrows.

On this understanding, any downward path in the diagram corresponds to an ordering of the objects through which it passes. Now suppose that each of the n objects exhibits one of the ranks of each of a set of ranked variables. For example, if there are two variables, A, a dichotomy into male and female (ranked 1, 2), and B, a fourfold division into social class, working, lower-middle, upper-middle and upper (ranked 1, 2, 3, 4), any individual under consideration will bear one of the ranks for each of A and B; the ranking (23), for instance, denoting a female in the upper middle class. In this particular case there are $2 \times 4 = 8$ types of individual and the number n of objects is accordingly 8.

If comparisons are made between neighbouring members differing only in one variable, beginning with (1, 1) and ending with (2, 4) we have what is called a partial ordering. Thus (11), (12), (13), (23), (24) is such an ordering, and so is (11), (21), (22), (23), (24). The situation is illustrated in Fig. 1, both partial orderings corresponding to downward paths on the diagram. Generally, if the objects correspond to rankings of $p_1, p_2, \ldots, p_r$ a partial ordering will be a set of $p_1 + p_2 + \ldots + p_r - r + 1$ objects such that any consecutive pair is of the type $(a, b, \ldots, j, \ldots, r)$, $(a, b, \ldots, j+1, \ldots, r)$. The number n is equal to $p_1 p_2 \ldots p_r$.

There will be $n!$ rankings of the n objects and $\dfrac{(\Sigma p_j - r)!}{(p_1 - 1)!\,(p_2 - 1)! \ldots (p_r - 1)!}$ partial orderings beginning with (1 1 1 ... 1) and ending with $(p_1 p_2 p_3 \ldots p_r)$. For in the process of following through any partial ordering any rank, say the jth, has to make $p_j - 1$ unit moves and these can occur at any point in the $\Sigma(p_j - 1)$ moves. Thus with $p_1 = 2, p_2 = 4$ there are four partial orderings:

$$\left.\begin{array}{lllll} (11) & (21) & (22) & (23) & (24) \\ (11) & (12) & (22) & (23) & (24) \\ (11) & (12) & (13) & (23) & (24) \\ (11) & (12) & (13) & (14) & (24) \end{array}\right\} \quad (1)$$

With $p_1 = 2$, $p_2 = 2$, $p_3 = 2$ there are six partial orderings each comprising four objects; and so on. The totality of partial orderings may be called the set of such orderings.

Consider a ranking of the n objects, e.g. for 2×4 objects,

$$(11)\ (12)\ (13)\ (14)\ (21)\ (22)\ (23)\ (24). \qquad (2)$$

We see that every pair of objects in the partial orderings (1) are in the same order as in the ranking (1). Such a ranking is said to be *consistent* with the set of partial orderings. It is not unique, e.g.

$$(11)\ (12)\ (13)\ (21)\ (14)\ (22)\ (23)\ (24) \qquad (3)$$

has the same property. Any other ranking may be compared with one of the consistent rankings by considering the minimum number of interchanges of adjacent pairs necessary to transform one to the other. More than one consistent ranking may be reached by the same minimum number of interchanges.

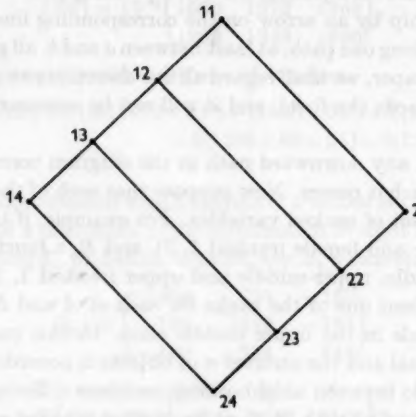


Fig. 1. Partially ordered system for two variables, one dichotomous and the other fourfold.

The problem discussed in this paper may be formulated as follows: A set of partial orderings is given, and a ranking of the objects is observed. Can this be regarded as consistent with the partial orderings? If the observed ranking is not subject to error, of course, this question is simply decided by examining whether it is one of the consistent rankings. We may also, however, allow for error in the observed ranking and regard it as consistent in a probabilistic sense if it is close to a consistent ranking, 'closeness' in this sense being measured by the minimum number of interchanges of adjacent pairs necessary to transform the observed to a consistent ranking.

This number of interchanges will be denoted by s, and our object is to find the frequency distribution of s in the population of $n!$ possible rankings of the objects. Thus, the hypothesis that an observed ranking is consistent with a set of partial orderings will be accepted if s is small; or, equivalently, if the probability of the observed s or a lower value is below some assigned significance level. The procedure is similar to Kendall's use of s to measure the agreement between two rankings (cf. Kendall, 1955). The difference in our case is that

we require the agreement between an observed ranking and one of a number of possible consistent rankings, not a unique one. Kendall's τ is defined in terms of s by

$$\tau = 1 - \frac{4s}{n(n-1)}, \tag{4}$$

but for our purposes it will be sufficient, and more convenient, to work with s itself.

DISTRIBUTION OF s FOR THREE DICHOTOMOUS VARIABLES

I consider in the first place three dichotomous variables ($p_1 = p_2 = p_3 = 2$). There are, as noted above, six partial orderings of four. The corresponding diagram is given in Fig. 2. The 8 objects give rise to $8! = 40{,}320$ rankings. Of these 48 are consistent.

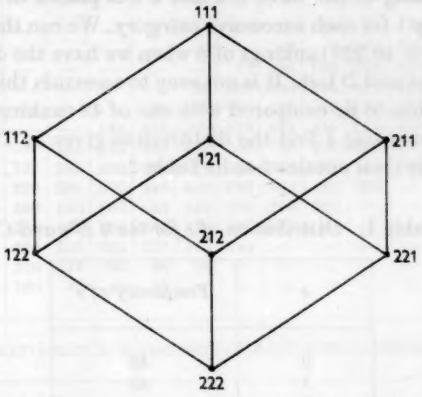


Fig. 2. Partially ordered system for three dichotomous variables.

This last result may be seen as follows: Since (111) comes first and (222) last we need only consider the other six objects. The six partial orderings are (112) (122); (112, 212); (121, 122); (121, 221); (211, 221); (211, 212). These do not impose any conditions on the order of the three (112), (121), (211) or of (122, 212, 221). But any member of the first precedes two members of the second. The first three must then occupy three of the first four places. It will be seen that if there is no overlap between members of the two sets the number of arrangements is $3! \times 3! = 36$. The only possible overlap occurs when one member of the first triad follows a member of the second in the fifth place. This gives rise to 12 further possibilities making 48 in all.

Let (111) be represented by A; each of (211), (121), (112) by B; each of (221), (212), (122) by C; and (222) by D.

In a ranking of eight A and D can occur in 56 different ways, which we can classify into 14 categories as follows:

A		—	A		—	—	A	
—		A	—		—	A	—	
—		—	—		A	—	—	
—	(1)	—	—	(2)	—	—	—	(3)
—		—	—		—	—	—	
—		—	—		—	—	D	
—		—	D		—	D	—	
D		D	—		D	—	—	

and so on up to

$$\begin{array}{ccccccc} - & - & - & - & - & - & A \\ - & - & - & - & - & A & D \\ - & - & - & - & A & D & - \\ - & - & - & A & D & - & - \\ - & - & A & D & - & - & - \\ - & A & D & - & - & - & - \\ A & D & - & - & - & - & - \\ D & - & - & - & - & - & - \end{array} \qquad (7)$$

The seven categories which follow are a horizontal mirror image of the first seven.

When any given ranking of the three B's and C's is placed in the blank positions it is found that s increases by 1 for each successive category. We can then build up the distribution of s in the $56 \times 720 = 40{,}320$ rankings of 8 when we have the distribution for one case, e.g. the array with A first and D last. It is not easy to ascertain this distribution. There are 720 members and each has to be compared with one of 48 rankings. By writing down the 720 permutations and counting I find the distribution given in Table 1. The distribution for the 40,320 rankings is then obtained as in Table 2.

Table 1. *Distribution of s for the* 6 *B's and C's*

s	Frequency of s
0	48
1	48
2	84
3	108
4	108
5	108
6	96
7	60
8	36
9	24
Total	720

Some further distributions for the case of two variates are given in Table 3 for the 2×2, 2×3, 2×4 and 3×3 cases.

With four dichotomous variables there are 16! rankings to be considered. (1111) and (2222) occur in 240 ways in a ranking of 16 and, as before, may be treated independently of the remaining 14 combinations in any enumeration for the distribution of s. Since direct enumeration of the distribution of s for the remaining 14 combinations would have been too tedious, an estimate of the form of the distribution was obtained by sampling.

The 14 combinations were written on slips which were placed in a bowl from which drawings were made to obtain a sample of some 2500 rankings. This number was held to be substantial in view of the 240 rankings of 16 associated with each of the 2500. The 240×2500

values of s were weighted according to the theoretical frequencies of certain sub-distributions to which they belonged. When graphed as a frequency polygon, the final estimate of the form of the distribution appeared almost identical with a normal curve. I am unable at this stage to give a more comprehensive account of the distributions of s and partially ordered systems.

Table 2. *Computations used to obtain the distribution of s for the rankings of the eight combinations of three dichotomous variables*

Category	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
1	—	—	—	—	—	—	—	—	—	—	—	—	—	24	36	60	96	108	108	108	84	48	48
2	—	—	—	—	—	—	—	—	—	—	—	—	48	72	120	192	216	216	216	168	96	96	—
3	—	—	—	—	—	—	—	—	—	—	—	72	108	180	288	324	324	324	252	144	144	—	—
4	—	—	—	—	—	—	—	—	—	—	96	144	240	384	432	432	432	336	192	192	—	—	—
5	—	—	—	—	—	—	—	—	—	120	180	300	480	540	540	540	420	240	240	—	—	—	—
6	—	—	—	—	—	—	—	—	144	216	360	576	648	648	648	504	288	288	—	—	—	—	—
7	—	—	—	—	—	—	—	168	252	420	672	756	756	756	588	336	336	—	—	—	—	—	—
8	—	—	—	—	—	—	168	252	420	672	756	756	756	588	336	336	—	—	—	—	—	—	—
9	—	—	—	—	—	144	216	360	576	648	648	648	504	288	288	—	—	—	—	—	—	—	—
10	—	—	—	—	120	180	300	480	540	540	540	420	240	240	—	—	—	—	—	—	—	—	—
11	—	—	—	96	144	240	384	432	432	432	336	192	192	—	—	—	—	—	—	—	—	—	—
12	—	—	72	108	180	288	324	324	324	252	144	144	—	—	—	—	—	—	—	—	—	—	—
13	—	48	72	120	192	216	216	216	168	96	96	—	—	—	—	—	—	—	—	—	—	—	—
14	24	36	60	96	108	108	108	84	48	48	—	—	—	—	—	—	—	—	—	—	—	—	—
Totals	24	84	204	420	744	1176	1716	2316	2904	3444	3828	4008	3972	3720	3276	2724	2112	1512	1008	612	324	144	48

Grand total, 40,320

MISSING COMBINATIONS

In the preceding, distributions of s are given for the situation where all of the possible combinations of the ranks of a set of variables are considered. It happens in practice, however, that some combinations do not appear at all or else appear with negligible frequencies. In such cases those combinations cannot be ranked with respect to the others. Since the problem discussed in this paper is to test if an observed ranking of combinations is consistent with a partially ordered system of the same combinations, we have to derive distributions based on partially ordered systems in which those combinations do not appear. This was done in connexion with the derivation of the distribution of s for the case of three dichotomous variables. Table 1 gives the distribution which is based on the partially ordered system in which both (111) and (222) are missing.

When combinations are considered a few at a time, the distributions so formed often fall into classes where within each class all are identical. For instance, when, in the case of three dichotomous variables, the combinations are considered 7 at a time we get either of two forms of distribution of s:

(a) Form F, if one of the B's or C's is missing; and

(b) Form G, if either A or D is missing.

The two forms of distribution are given in Table 4.

Table 3. *Two variable cases*

A, both dichotomous. B, one dichotomous and the other having three ranks.
C, one dichotomous and the other having four ranks. D, both having three ranks.

s	Frequencies			
	A	B	C	D
0	2	5	14	42
1	4	15	56	170
2	6	32	150	464
3	6	53	318	1,016
4	4	77	587	1,946
5	2	96	963	3,362
6	—	105	1,448	5,350
7	—	102	2,009	7,942
8	—	88	2,599	11,096
9	—	67	3,151	14,664
10	—	43	3,601	18,442
11	—	24	3,883	22,138
12	—	10	3,958	25,434
13	—	3	3,817	28,010
14	—	—	3,477	29,604
15	—	—	2,990	30,054
16	—	—	2,418	29,306
17	—	—	1,832	27,454
18	—	—	1,289	24,684
19	—	—	839	21,276
20	—	—	495	17,548
21	—	—	260	13,818
22	—	—	115	10,354
23	—	—	41	7,354
24	—	—	9	4,916
25	—	—	1	3,072
26	—	—	—	1,770
27	—	—	—	930
28	—	—	—	432
29	—	—	—	170
30	—	—	—	52
31	—	—	—	10
Totals	24	720	40,320	362,880

Table 4. *Distribution of s for three dichotomous variables with one combination missing*

s	Frequencies	
	Form F (6 distributions)	Form G (2 distributions)
0	16	48
1	48	96
2	104	180
3	184	288
4	284	396
5	396	504
6	500	600
7	572	612
8	604	600
9	588	540
10	528	432
11	436	324
12	324	216
13	220	120
14	132	60
15	68	24
16	28	—
17	8	—
Totals	5040	5040

Table 5. *Distribution of s for the case of two variables, one dichotomous and one fourfold, with one combination missing*

s	Frequencies (two distributions in each form)			
	Form P	Form Q	Form R	Form S
0	14	5	9	7
1	42	20	32	26
2	94	52	75	64
3	168	105	142	125
4	269	182	232	210
5	376	278	338	312
6	485	383	445	419
7	561	480	533	512
8	604	553	587	575
9	594	588	597	596
10	544	578	560	570
11	450	525	484	503
12	344	439	382	407
13	235	337	275	301
14	145	235	178	201
15	74	147	101	119
16	32	80	48	60
17	8	37	18	25
18	1	13	4	7
19	—	3	—	1
Totals	5040	5040	5040	5040

Again, when in the case of two variables where one is dichotomous and the other has four ranked classifications, the combinations are considered 7 at a time and we get four forms of distribution of s:

(a) Form P, if either (11) or (24) is missing;
(b) Form Q, if either (21) or (14) is missing;
(c) Form R, if either (12) or (23) is missing; and
(d) Form S, if either (13) or (22) is missing.

The four forms of distribution are given in Table 5.

The writer is indebted to Prof. F. Stuart Chapin for encouragement and to Prof. Palmer O. Johnson for guidance.

REFERENCE

KENDALL, M. G. (1955). *Rank Correlation Methods*, 2nd ed. London. Chas. Griffin and Co.

ON A CLASS OF SKEW DISTRIBUTION FUNCTIONS

By HERBERT A. SIMON†

Carnegie Institute of Technology

I. Introduction

It is the purpose of this paper to analyse a class of distribution functions that appears in a wide range of empirical data—particularly data describing sociological, biological and economic phenomena. Its appearance is so frequent, and the phenomena in which it appears so diverse, that one is led to the conjecture that if these phenomena have any property in common it can only be a similarity in the structure of the underlying probability mechanisms. The empirical distributions to which we shall refer specifically are: (A) distributions of words in prose samples by their frequency of occurrence, (B) distributions of scientists by number of papers published, (C) distributions of cities by population, (D) distributions of incomes by size, and (E) distributions of biological genera by number of species.

No one supposes that there is any connexion between horse-kicks suffered by soldiers in the German army and blood cells on a microscope slide other than that the same urn scheme provides a satisfactory abstract model of both phenomena. It is in the same direction that we shall look for an explanation of the observed close similarities among the five classes of distributions listed above.

The observed distributions have the following characteristics in common:

(*a*) They are J-shaped, or at least highly skewed, with very long upper tails. The tails can generally be approximated closely by a function of the form

$$f(i) = (a/i^k)\,b^i, \tag{1·1}$$

where a, b, and k are constants; and where b is so close to unity that in first approximation the final factor has a significant effect on $f(i)$ only for very large values of i. Thus, for example, the number of words that occur exactly i times in James Joyce's *Ulysses* is about a/i^k; the number of authors who published exactly i papers in *Econometrica* over a twenty-year period is approximately a/i^k; and so on.

(*b*) The exponent, k, is greater than 1, and in the cases of word frequencies, publication, and urban populations is very close to 2.‡

(*c*) In the cases of word frequencies, publications and biological genera, the function (1·1) describes the distribution not merely in the tail but also for small values of i. In these cases the ratio $f(2)/f(1)$ is generally in the neighbourhood of one-third, and almost never reaches one-half; while $f(1)/n$, where $n = \sum_1^\infty f(i)$, is generally in the neighbourhood of one-half.

Property (*a*) is characteristic of the 'contagious' distributions—for example, the negative binomial as it approaches its limiting form, Fisher's logarithmic series distribution. However, in the case of the negative binomial, k cannot exceed unity (and equals unity only in

† I have had the benefit of helpful comments from Messrs Benoit Mandelbrot, Robert Solow and C. B. Winsten. I am grateful to the Ford Foundation for a grant-in-aid that made the completion of this work possible.

‡ See Zipf (1949) for numerous examples of distributions with this property.

the limiting case of the log series); and if the distribution has a long tail, so that the convergence factor, b, is close to unity, $f(2)/f(1)$ cannot be less than one-half. Hence the negative binomial and Fisher's log series distributions do not provide a satisfactory fit for data possessing property (a) together with either (b) or (c).†

It is well known that the negative binomial and the log series distributions can be obtained as the stationary solutions of certain stochastic processes. For example, J. H. Darwin (1953) derives these from birth and death processes, with appropriate assumptions as to the birth- and death-rates and the initial conditions. In this paper we shall show that stochastic processes closely similar to those yielding the negative binomial or log series distributions lead to a class of functions having the three properties enumerated above. This class of functions is given by

$$f(i) = AB(i, \rho+1), \tag{1·2}$$

where A and ρ are constants, and $B(i, \rho+1)$ is the Beta function of $i, \rho+1$:

$$B(i,\rho+1) = \int_0^1 \lambda^{i-1}(1-\lambda)^\rho \, d\lambda = \frac{\Gamma(i)\,\Gamma(\rho+1)}{\Gamma(i+\rho+1)} \quad (0<i;\ 0<\rho<\infty). \tag{1·3}$$

Now it is a well-known property of the Gamma function (Titchmarsh, 1939, p. 58) that as $i \to \infty$, and for any constant, k,

$$\frac{\Gamma(i)}{\Gamma(i+k)} \sim i^{-k}. \tag{1·4}$$

Hence, from (1·3), we have, as $i \to \infty$:

$$f(i) \sim \Gamma(\rho+1)\, i^{-(\rho+1)}. \tag{1·5}$$

Therefore, the distribution (1·2) approximates the distribution (1·1) in the tail (more precisely, through the range in which the convergence factor of the latter is close to one). Further, if ρ is positive, k will be greater than 1, as required by (b); and if ρ is equal to 1, k will be equal to 2. It is easy to see that in the latter case we will have

$$f(i) = \frac{1}{i(i+1)}, \quad \sum_{i=1}^{\infty} f(i) = 1, \tag{1·6}$$

so that $f(2)/f(1) = \frac{1}{3}$; and $f(1)/n = \frac{1}{2}$, as required by (c).

In the remainder of this paper I propose: (a) to describe a stochastic process that leads to the stationary distribution (1·2); (b) to discuss some generalizations of this process; and (c) to construct hypotheses as to why the empirical phenomena mentioned above can be represented, approximately, by processes of this general kind. Before proceeding, I should like to mention two earlier derivations, one of (1·2), the other of (1·1), that I have been able to discover in the literature.

Some thirty years ago, G. Udny Yule (1924) constructed a probability model, with (1·2) as its limiting distribution, to explain the distribution of biological genera by numbers of species. He also derived a modified form of (1·2), replacing the complete Beta-function of (1·3) by the incomplete Beta-function with upper limit of integration $\alpha < 1$. (This modification has the same effect as the introduction of the convergence factor, b^i, in (1·1)—it causes a more rapid decrease in $f(i)$ for very large values of i; cf. also Darwin (1953, p. 378).) It seems highly appropriate to call the distribution (1·2) the Yule distribution.

† The contrasting characteristics of distributions for which the log series provides a satisfactory fit and those, under consideration here, for which it does not are illustrated by examples (i) and (ii), respectively, in Good (1953).

Because Yule's paper predated the modern theory of stochastic processes, his derivation was necessarily more involved than the one we shall employ here. Moreover, while the assumptions he required are plausible for the particular biological problem he treated, the corresponding assumptions applied to the four other phenomena we have mentioned appear much less plausible. Our derivation requires substantially weaker assumptions than Yule's about the underlying probability mechanism.

More recently D. G. Champernowne (1953) has contructed a stochastic model of income distribution that leads to (1·1) and to generalizations of that function. Since the points of similarity between his model and the one under discussion here are not entirely obvious at a first examination, I shall consider their relation in a later section of this paper.

II. THE STOCHASTIC MODEL

For ease of exposition, the model will be described in terms of word frequencies. In a later section, alternative interpretations will be provided. Our present interest is in the kind of stochastic process that would lead to (1·2).

Consider a book that is being written, and that has reached a length of k words. We designate by $f(i,k)$ the number of *different* words that have occurred exactly i times in the first k words. That is, if there are 407 different words that have occurred exactly once each, then $f(1,k) = 407$.

Assumption I. The probability that the $(k+1)$*-st word is a word that has already appeared exactly* i *times is proportional to* $if(i,k)$—that is, to the total number of occurrences of all the words that have appeared exactly i times.

Note that this assumption is much weaker than the assumption (I′): that the probability a *particular* word occur next be proportional to the number of its previous occurrences. Assumption (I′) implies (I), but the converse is not true. Hence we leave open the possibility that, among all words that have appeared i times the probability of recurrence of some may be much higher than of others.

Assumption II. There is a constant probability, α*, that the* $(k+1)$*-st word be a new word*—a word that has not occurred in the first k words.

Assumptions (I) and (II) describe a stochastic process, in which the probability that a particular word will be the next one written depends on what words have been written previously. If this process correctly describes the selection of words, then the words in a book cannot be regarded as a random sample drawn from a population with a prior distribution. The reasonableness of the former, as compared with the latter type of explanation of the observed distributions, will be discussed in § IV.

From (I), it follows that

$$\mathscr{E}\{f(i,k+1)\} - f(i,k) = K(k)\{(i-1)f(i-1,k) - if(i,k)\} \quad (i = 2,\ldots,k+1), \tag{2·1}$$

for if the $(k+1)$st word is one that has previously occurred $(i-1)$ times, $f(i,k+1)$ will be increased over $f(i,k)$, and the probability of this, by assumption (I), is proportional to $(i-1)f(i-1,k)$; if the $(k+1)$st word is one that previously occurred i times, $f(i,k+1)$ will be decreased, and the probability of this, by assumption (I), is proportional to $if(i,k)$; while in all other cases, $f(i,k+1) = f(i,k)$.

From (I) and (II) we obtain similarly

$$\mathscr{E}\{f(1,k+1)\} - f(1,k) = \alpha - K(k)f(1,k) \quad (0 < \alpha < 1). \tag{2·2}$$

Since we will be concerned throughout with 'steady-state' distributions (as defined by equation (2·8) below), we replace the expected values in (2·1) and (2·2) by the actual frequencies. (Alternatively, we might replace frequencies on the right-hand side of the equation by probabilities.) That is, we write, instead of (2·1) and (2·2),

$$f(i,k+1)-f(i,k) = K(k)\{(i-1)f(i-1,k)-if(i,k)\} \quad (i=2,\ldots,k+1), \tag{2·3}$$

$$f(1,k+1)-f(1,k) = \alpha - K(k)f(1,k), \tag{2·4}$$

where the f's now represent expected values.

Now, we wish to evaluate the factor of proportionality $K(k)$. Since $K(k)if(i,k)$ is the probability that the $(k+1)$st word is one that previously occurred i times, we must have

$$\sum_{i=1}^{k} K(k)if(i,k) = K(k)\sum_{i=1}^{k} if(i,k) = 1-\alpha. \tag{2·5}$$

But $\sum_{i=1}^{k} if(i,k)$ is the total number of words up to the kth, hence

$$\sum_{i=1}^{k} if(i,k) = k, \tag{2·6}$$

and

$$K(k) = \frac{1-\alpha}{k}. \tag{2·7}$$

Substituting (2·7) in (2·3) and (2·4), we could solve these differential equations explicitly. We can avail ourselves, however, of a simpler—though non-rigorous—method for discovering the solutions, and can then test their correctness by substitution in the original equations. Consider the 'steady-state' distribution in the following sense. We assume

$$\frac{f(i,k+1)}{f(i,k)} = \frac{k+1}{k} \quad \text{for all } i \text{ and } k; \tag{2·8}$$

so that all the frequencies grow proportionately with k, and hence maintain the same relative size. (Since we must have $f(i,k)=0$ for $i>k$, equation (2·8) cannot hold exactly for all i and k. But as explained above, we are concerned at the moment with heuristic rather than proof.)

From (2·8) it follows that

$$\frac{f(i,k)}{f(i-1,k)} = \frac{f(i,k+1)}{f(i-1,k+1)} = \beta(i), \tag{2·9}$$

where $\beta(i)$ does not involve k. Hence, the *relative* frequencies, which we will designate by $f^*(i)$, are independent of k. Substituting (2·7), (2·8) and (2·9) in (2·3), we get

$$\left(\frac{k+1}{k}-1\right)f(i,k) = \frac{(1-\alpha)}{k}\left\{\frac{(i-1)}{\beta(i)}-i\right\}f(i,k). \tag{2·10}$$

Cancelling the common factor, and solving β for (i), we obtain

$$\beta(i) = \frac{(1-\alpha)(i-1)}{1+(1-\alpha)i} = \frac{f^*(i)}{f^*(i-1)} \quad (i=2,\ldots,k). \tag{2·11}$$

For convenience, we introduce

$$\rho = \frac{1}{1-\alpha} \quad (1<\rho<\infty). \tag{2·12}$$

Since $f^*(i) = \beta(i) f^*(i-1) = \beta(i).\beta(i-1)\dots\beta(2) f^*(1)$, we obtain from (2·11) and (2·12)

$$f^*(i) = \frac{(i-1)(i-2)\dots 2.1}{(i+\rho)(i+\rho-1)\dots(1+\rho)} f^*(1) = \frac{\Gamma(i)\,\Gamma(\rho+1)}{\Gamma(i+\rho+1)} f^*(1) = B(i,\rho+1) f^*(1) \quad (i = 2,\dots,k). \tag{2·13}$$

The second relation follows from the fact that

$$\Gamma(i+\rho+1) = (i+\rho)\,\Gamma(i+\rho) = (i+\rho)(i+\rho-1)\dots(1+\rho)\,\Gamma(\rho+1). \tag{2·14}$$

But (2·13) is identical with (1·2) if we take $A = f^*(1)$.

That (2·13) is in fact a solution of (2·3) can be verified by direct substitution. Moreover, it is in the following sense a stable solution. Suppose that (2·11) is *not* satisfied. Whatever be the values of the $f(i,k)$ for a given k, we may write without loss of generality

$$\frac{f(i,k)}{f(i-1,k)} = \frac{(1-\alpha)(i-1)}{(1-\alpha)i + 1 + \epsilon(i,k)}, \tag{2·15}$$

where $\epsilon(i,k)$ is some function of i and k. If we now divide both sides of (2·3) by $f(i,k)$ and substitute (2·15) in the right-hand side of the resulting equation, we obtain after simplification

$$\frac{f(i,k+1)}{f(i,k)} = \frac{k+1+\epsilon(i,k)}{k}. \tag{2·16}$$

Hence the ratio of $f(i,k+1)$ to $f(i,k)$ will be greater than $(k+1)/k$ if $\epsilon(i,k)$ is positive, and less than $(k+1)/k$ if $\epsilon(i,k)$ is negative. Since new words are introduced at a constant rate, $\sum_1^k f(i,k)$ must be proportional to k; therefore, by (2·16), we will have

$$\sum_{i=1}^{k+1} f(i,k+1) - \frac{k+1}{k} \sum_{i=1}^{k+1} f(i,k) = \frac{1}{k} \sum_{i=1}^{k} \epsilon(i,k) f(i,k) = 0. \tag{2·17}$$

We may interpret the three equations, (2·15)–(2·17), as follows. In an average sense, the frequencies will grow proportionately with k. If a particular frequency is 'too large' compared with the next lower frequency ($\epsilon(i,k)$ negative in (2·15)), it will grow at a rate slower than the average; if it is 'too small' ($\epsilon(i,k)$ positive), it will grow more rapidly than the average.

It remains to be shown that $f^*(i) = B(i,\rho+1) f^*(1)$ is a proper distribution function. In particular, we require that $\sum_{i=1}^{k} iB(i,\rho+1)$ converge as $k \to \infty$. Now, it is well known that $\sum_{i=1}^{\infty} i^{-a}$ converges for every $a > 1$. But by (1·4),

$$\sum_{i=1}^{\infty} iB(i,\rho+1) \sim i.i^{-(\rho+1)} = i^{-\rho}. \tag{2·18}$$

Hence, by the usual ratio comparison test, $\sum_{i=1}^{\infty} iB(i,\rho+1)$ converges for $\rho > 1$, as required.

From the definition of α the total number, n_k, of *different* words will be αk; while the total number of word occurrences is k. That is

$$n_k = \sum_{i=1}^{k} f(i,k) = \alpha k = \alpha \sum_{i=1}^{k} i f(i,k). \tag{2·19}$$

Returning to (2·4), and using (2·8), we get

$$\left(\frac{k+1}{k}-1\right)f^*(1)=\alpha-\frac{1-\alpha}{k}f^*(1), \tag{2·20}$$

whence

$$f^*(1)=\frac{k\alpha}{2-\alpha}=\frac{n_k}{2-\alpha}. \tag{2·21}$$

From (2·12) and (2·21), and by successive application of (2·11), we can compute the values of $\rho, f^*(1)/n_k, f^*(2)/n_k, f^*(3)/n_k$, etc., for given values of α (Table 1).

Table 1

α	ρ	$f^*(1)/n_k$	$f^*(2)/n_k$	$f^*(3)/n_k$
0·0	1	0·500	0·167	0·083
0·1	1·11	0·527	0·169	0·082
0·2	1·25	0·556	0·171	0·080
0·3	1·43	0·588	0·171	0·077
0·5	2·00	0·667	0·167	0·067
0·7	3·33	0·769	0·144	0·046
0·9	10·00	0·909	0·076	0·012

Thus far we have considered the case where α, the rate at which new words are introduced, is independent of k. We can easily generalize to the case where α is a function of k by making the appropriate substitution in (2·4). The equations can then be solved directly, but the method employed to obtain a 'steady-state' distribution is not applicable, since it is not easy to define what is meant by the steady state in this more general case. We will content ourselves with some approximate results for two special cases. These special cases will give us insight as to how a distribution function may arise which, for small values of i, can be approximated by (1·2), with $0<\rho<1$.

Case I. Suppose the system to be in the steady state described by (2·13) with $k=k_0$, and that the flow of *new* words suddenly ceases, so that $\alpha(k)=0$ for $k>k_0$. We will now have $K(k)=1/k$ for $k>k_0$, and (2·4) becomes

$$f(1,k+1)=\left(1-\frac{1}{k}\right)f(1,k)=\frac{k-1}{k}f(1,k). \tag{2·22}$$

We define

$$\gamma(i)=\frac{f(i,k+1)}{f(i,k)} \quad (i=2,\ldots,k+1). \tag{2·23}$$

Since no new words are being introduced, we must have

$$\begin{aligned} n_k &= f(1,k)+\sum_{i=2}^{k}f(i,k)=f(1,k+1)+\sum_{i=2}^{k}f(i,k+1)\\ &=\frac{(k-1)}{k}f(1,k)+\sum_{i=2}^{k}\gamma(i)f(i,k), \end{aligned} \tag{2·24}$$

whence

$$\frac{\sum_{i=2}^{k}[\gamma(i)-1]f(i,k)}{\sum_{i=2}^{k}f(i,k)}=\frac{1}{k}\frac{f(1,k)}{\sum_{i=2}^{k}f(i,k)}. \tag{2·25}$$

Let us define next

$$\beta(i) = \frac{f(i,k)}{f(i-1,k)} = \frac{(i-1)}{(1+\rho_i)} \tag{2·26}$$

(where we suppose that ρ_i changes only slowly with k). Instead of (2·3), we have

$$f(i,k+1) - f(i,k) = \frac{1}{k}[(i-1)f(i-1,k) - if(i,k)]. \tag{2·27}$$

Substituting (2·23) and (2·26) in this, we get

$$\gamma(i) - 1 = \frac{1}{k}[(i+\rho_i) - i], \tag{2·28}$$

whence

$$\rho_i = k(\gamma(i) - 1), \tag{2·29}$$

and

$$\bar{\rho} = \frac{\sum_{i=2}^{k} k(\gamma(i)-1)f(i,k)}{\sum_{i=2}^{k} f(i,k)} = \frac{f(1,k)}{\sum_{i=2}^{k} f(i,k)} = \frac{f(1,k)}{n_k - f(1,k)}. \tag{2·30}$$

Define

$$p_1 = f(1,k)/n_k. \tag{2·31}$$

Then

$$\bar{\rho} = \frac{p_1}{1-p_1} \quad \text{and} \quad 0 < \bar{\rho} < \infty. \tag{2·32}$$

Proceeding heuristically, we can see that after α becomes zero, $f(1,k)$ will begin to decrease with k, and the value of ρ_i will be larger the larger is i. For small values of i, we will have $\rho(i) < \bar{\rho}$, and for large values, $\rho(i) > \bar{\rho}$. However, the tail of the distribution will be affected only slowly by the change in α. Hence, we may suppose that $\lim_{i \to k_0} \rho(i) = \rho_0$, where ρ_0 is $\rho(k_0)$. On the other hand, since the weighted average in (2·29) is heavily influenced by the large frequencies for small values of i, ρ_i will be only slightly less than $\bar{\rho}$. Hence we may expect the distribution to take the form of a slightly curved line on a double-log scale, with a slope of $-(\bar{\rho}+1)$ at the lower end, and a slope of $-(\rho_0+1)$ at the upper end. If $\rho_0 > 2$, then $\Sigma i f(i,k)$ will converge. An example of such a distribution will be given in § IV.

Case II. A second approximate solution can be obtained if we assume that α decreases with k, but very slowly. By definition, we have $\alpha = dn_k/dk = n'$. The condition for a steady state (all frequencies increasing proportionately) is now

$$f(i,k+1) = [1 + (n'/n_k)]f(i,k). \tag{2·33}$$

Substituting as before, (2·7) and (2·33) in (2·3), we again obtain (2·13), where ρ is now given by

$$\rho = \frac{n'k}{n_k} \frac{1}{(1-n')}. \tag{2·34}$$

The slope obtained in the derivation for constant α has now been multiplied by the factor $(n'k)/n_k$, which for monotonically decreasing α is less than one. Hence, the effect of a decrease in the rate of introduction of new words is to lengthen the tail of the distribution, as was also true in case I. If the new value of ρ is less than one, we do not have a proper distribution function (see equation (2·18)), hence the equation can hold only for small and moderate values of i, and there must be a curve (on a logarithmic scale) in the tail of the distribution.

III. An alternative formulation of the process

There are some alternative ways for deriving the relation (2·13). One of these will be useful to us when we come, in the next section, to a more specific discussion of word frequencies and frequencies of publications. Moreover, this derivation avoids the difficulties we have encountered in the definition of 'steady state'.

Equation (2·10) may be written

$$0 = (1-\alpha)[(i-1)f^*(i-1) - if^*(i)] - f^*(i) \quad (i = 2, \ldots, k), \tag{3·1}$$

where we have again written $f^*(i)$ for $f(i, k)$.

Similarly, from (2·4), we obtain

$$0 = 1 - (1-\alpha)f^*(1) - f^*(1). \tag{3·2}$$

These two equations may be interpreted as follows. We consider a sequence of k words. We add words to the sequence in accordance with assumptions (I) and (II) of § II, *but we drop words from the sequence at the same average rate*, so that the length of the sequence remains k. The method according to which we drop words is the following:

Assumption III. If one representative of a particular word is dropped, then all representatives of that word are dropped, and the probability that the next word dropped be one with exactly i representatives is $f^*(i)$.

This assumption would be approximately satisfied, for example, if the representatives of each word, instead of being distributed randomly through the sequence, were closely 'bunched'. This possibility is consistent with assumption (I).

Equation (3·1), in our new interpretation, may be regarded as the steady-state equilibrium of the stochastic process defined by

$$f(i, m+1) - f(i, m) = (1-\alpha)[(i-1)f(i-1, m) - if(i, m)] - f(i, m), \tag{3·3}$$

where m is now not the total number of words (which remains a constant, k), but the number of additions to (and withdrawals from) an initial arbitrary sequence of k words. Since the k of this process, unlike that of § II, remains constant, the ordinary proofs of the existence of a unique steady-state solution will apply (see Feller, 1950, p. 373), and we avoid the troublesome questions of rigour that confronted us in § II.

The solution of (3·1) and (3·2) is, of course, again given by

$$\frac{f^*(i)}{f^*(i-1)} = \frac{(1-\alpha)(i-1)}{1+(1-\alpha)i}. \tag{2·11}$$

If we were to replace the last terms of (3·1) and of (3·2), respectively, by terms corresponding to the usual form of the death process, we would have (cf. Darwin, 1953, p. 375; and Kendall, 1948)

$$0 = (1-\alpha)[(i-1)f^*(i-1) - if^*(i)] - [if^*(i) - (i+1)f^*(i+1)] \quad (i = 2, \ldots, k-1), \tag{3·4}$$

$$0 = 1 - (1-\alpha)f^*(1) - [f^*(1) - 2f^*(2)]. \tag{3·5}$$

The solution of this system of equations is

$$\frac{f^*(i)}{f^*(i-1)} = \frac{(1-\alpha)(i-1)}{i}, \tag{3·6}$$

which is Fisher's logarithmic series distribution.

Since the log series distribution is a limiting case of the negative binomial, we may ask whether there is a distribution that stands in the same relation to the latter as (2·11) stands in relation to (3·6). We can obtain such a distribution by a modification of the birth process in (3·1). We assume now that the birth-rate is the sum of two components—one proportional to $if(i)$, the other proportional to $f(i)$. In place of (3·1) we have

$$0 = \frac{(1-\alpha)k}{k+c}[(i-1+c)f^*(i-1)-(i+c)f^*(i)]-f^*(i) \quad (c \text{ a constant}), \tag{3·7}$$

the solution of which is

$$\frac{f^*(i)}{f^*(i-1)} = \frac{\lambda(i-1+c)}{\lambda(i+c)+1} = \frac{(i-1+c)}{(i+c+1/\lambda)}, \tag{3·8}$$

where

$$\lambda = k(1-\alpha)/(k+c).$$

A rather remarkable property of (3·8) is that in the tail it still has the limiting form (1·1) with $b = 1$. Hence for α and c small, this generalized Yule distribution will still possess the three properties listed in the introduction. The fact that a reasonably wide range of variation in the assumptions underlying the stochastic model does not alter greatly the form of the distribution adds plausibility to the use of such stochastic processes to explain the observed distributions. Our next task is to consider these explanations in more detail.

IV. The empirical distributions

In this section I shall try to give theoretical justifications for the observed fit of the Yule distribution to a number of different sets of empirical data.

A. *Word frequencies*

A substantial number of word counts have been made, in English and in other languages (see Hanley, 1937; Thorndike, 1937; Yule, 1944; Zipf, 1949; and Good, 1953). Equation (1·6) provides a good fit to almost all of them. When the more general function, (1·2), is used, the estimated value of ρ is always close to 1. When a convergence factor, b^i, is introduced to account for the deficiency in frequencies for very large values of i, the estimated value of b is also very close to 1. Good (1953), for instance, applies (1·6) multiplied by a convergence factor to the Eldridge count, and obtains $b = 0{\cdot}999667$.

These regularities are the more surprising in that the various counts refer to a quite heterogeneous set of objects. In the Yule and Thorndike counts, inflected forms are counted with the root word; in most of the other counts each form is regarded as a distinct word. The Yule counts include only nouns; the others, all parts of speech. The Dewey, Eldridge and Thorndike counts are composite—compiled from a large number of separate writings; most of the others are based on a single piece of continuous prose. I would regard this heterogeneity as further evidence that the explanation is to be sought in a probability mechanism, rather than in more specific properties of language; but at the same time, the heterogeneity complicates the task of specifying the probability mechanism in detail. I shall avoid questions of 'fine structure'—which would require an expertness in linguistics that I do not possess—and confine myself to three broad problems: (1) the distribution of word frequencies in the whole historical sequence of words that constitutes a language; (2) the distribution of word frequencies in a continuous piece of prose; (3) the distribution of word frequencies in a sample of prose assembled from composite sources.

(1) For obvious reasons, we do not have any empirical data on the cumulated word frequencies for a whole language. On *a priori* grounds, it does not appear unreasonable to postulate that these frequencies are determined by a process like that described in § II. The parameter α is then the rate at which neologisms appear in the language as a fraction of all word occurrences—and hence α can be assumed to be very close to zero.

(2) The process of § II might also describe the growth of a continuous piece of prose—for example, Joyce's *Ulysses*. But there are some serious objections to this hypothesis. An author writes not only by processes of *association*—i.e. sampling earlier segments of the word sequence—but also by processes of *imitation*—i.e. sampling segments of word sequences from other works he has written, from works of other authors, and, of course, from sequences he has heard. The model of §II apparently allows only for association, and excludes imitation.

The word frequencies in *Ulysses* provide obvious evidence of the importance of both processes. The fact that the proper noun 'Bloom' occurs 926 times and ranks 30th in frequency must be attributed to association. If Joyce had named his hero 'Smith', that noun, instead of 'Bloom', would have ranked 30th. On the other hand, 'they', which occurs 1010 times in *Ulysses* and ranks 27th, has very nearly the same rank—the 28th—in the Dewey count. In fact, of the 100 most frequent words in *Ulysses*, 78 are among the top 100 in the Dewey count. This similarity in ranking of 'common' words argues for imitation rather than association. Even for the common words, however, the variations in frequency from one count to another are far too great to be explained as fluctuations resulting from random sampling from a common population of words. The imitative process must involve stratified sampling, and imitation must be compounded with association.

It is worth emphasizing again at this point that assumption (I) does not require that the choice of the next word from among those previously written be completely random. Suppose, for example, that a writer were to assign to each page he has already written a number, p_j, $\Sigma p_j = (1-\alpha)$, the size of p_j varying with the 'affinity' of the subject discussed on the jth page to the subject next to be discussed. If his next word were selected by a stratified sampling of the previous pages, with probability p_j for each page, then equation (2·1) would generally be satisfied. For although individual words would be distributed unevenly through the preceding pages, the totality of words having a given frequency, i, in all the previous pages taken together would be distributed almost evenly through these pages. Hence, the various frequency strata would have proportionate probabilities of being sampled, for most choices of the p_j. This is all that is required for equation (2·1). This same comment applies to the assumption we shall subsequently make regarding imitative sampling from other works.

Let us now reconsider the problem of a piece of continuous prose. Since both the processes of association and imitation are involved, the sequence that is counted is to be regarded as a 'slice', of length k, of the entire sequence of words in the language, or of the entire sequence written by the author. Hence the word count is better described by the stochastic process of § III than by the process of § II.

In determining the probability that a word selected in such a sequence be one that has occurred exactly i times, we must consider separately the process of imitation and association. Assume that, on the average, a fraction, β, of the words added is selected by imitation, and the remaining fraction, $(1-\beta)$, by association. Since no new words can be introduced by association, the joint probability that the next word will be selected by association and will be a word that has already occurred i times is $(1-\beta)if(i,k)/k$.

The words selected by imitation present a more difficult problem, and we shall have to content ourselves with a reasonable assumption that has no rigorous justification. On the average, a word that has occurred i times will have a chance less than i/k of being the next one chosen by imitation, because in the sequence that is being sampled there are words that have not yet been chosen at all, and because with progressive change of subject, different strata of the language will be sampled. Since words with large i will generally be 'common' words, fairly uniformly distributed through all strata of the language, the deficiency may be expected to be proportionately greater for small i than for large i. As a rough, but reasonable, approximation let us assume that: the joint probability that the next word will be selected by imitation and will be a word that has already occurred i times is $\beta(i-c)f(i,k)/k$, where $0<c<1$. (Our result would not be essentially altered if we wrote $c(i)$ instead of c, provided only that $c(i)$ does not vary a great deal.)

Adding the two joint probabilities—for association and imitation, respectively—we find that the total probability that the next word be one that has occurred i times is $(i-\beta c)f(i,k)/k$. By summing this probability over i and subtracting from 1, we find that the probability that the next word be a new word is $\beta c(n_k/k)$.

If the method of dropping words from the sequence satisfies assumption (III) of § III, we set the difference between the birth-rate and the death-rate equal to zero, and obtain the steady-state equation

$$0=(i-c\beta-1)f^*(i-1)-(i-c\beta)f^*(i)-f^*(i), \tag{4·1}$$

which has as its solution

$$\frac{f^*(i)}{f^*(i-1)}=\frac{(i-c\beta-1)}{(i-c\beta+1)}. \tag{4·2}$$

Again, we obtain a distribution with the required properties.

(3) The distribution of word frequencies in a sample of prose assembled from composite sources can be explained along the same general lines. Again, we may regard the sample as a 'slice' from a longer sequence, but we might expect the parameters c and β to be somewhat larger than in a comparable piece of continuous prose. The qualification 'comparable' is important, for c may be expected to be smaller for homogeneous prose using a limited vocabulary of common words than for prose with a large vocabulary and treating of a variety of subjects. Hence c might well be larger for the continuous *Ulysses* count than for the Eldridge count, which is drawn from newspaper sources. Indeed, the empirical evidence suggests that this is the case.

There is no point in elaborating the explanation further. What has been shown is that the observed frequencies can be fitted by distributions derived from probability assumptions that are not without plausibility.

A very different and very ingenious explanation of the observed word-frequency data has been advanced recently by Dr Benoit Mandelbrot (1953). His derivation rests on the assumption that the frequencies are determined so as to maximize the number of bits of information, in the sense of Shannon, transmitted per symbol. There are several reasons why I prefer an explanation that employs averaging rather than maximizing assumptions. First, an assumption that word usage satisfies some criterion of efficiency appears to be much stronger than the probability assumptions required here. Secondly, numerous doubts, which I share, have been expressed as to the relevance of Shannon's information measure for the measurement of semantic information.

Before leaving the subject of word frequencies, it may be of interest to look at some of the empirical data. Good (1953, pp. 257–60), has obtained good fits to the Eldridge count and to one of Yule's counts by the use of equation (1·6). Table 2, summarizes a few of the data on two word counts, and compares the actual frequencies, $f(1)$, $f(2)$ and $f(3)$ with the frequencies estimated from equation (1·3). The actual values of k and n_k are used to estimate $\alpha = n_k/k$, and (2·11) and (2·21) to obtain the expected frequencies. In both cases the observed value of n_k/k leads to an estimate of ρ in the neighbourhood of 1·1 to 1·2. An empirical fit to the whole distribution of a function of the form $f(i) = Ka^{-(\rho+1)}$ gives an estimated value of ρ, in both cases, of about one—in reasonable agreement with the first estimate. A good fit to both the *Ulysses* and the Eldridge counts can also be obtained from (4·2), with c equal to about 0·2 in the former case, and close to zero in the latter.

In the case of Thorndike's count of $4\frac{1}{2}$ million words in children's books (Thorndike, 1937), we may assume that the supply of new words was virtually exhausted before the end of the count. In his count $f(1)$ is substantially below $0{\cdot}5n_k$ (about $0{\cdot}34n_k$), as we would expect under these circumstances (see case I of § II). Thorndike estimated the empirical value of our $\overline{\rho}$ at 0·45, which is entirely consistent with the observed value of $0{\cdot}34n_k$ for $f(1)$. For, by (2·32), $p_1 = \overline{\rho}/(\overline{\rho}+1) = 0{\cdot}31$.

Table 2

Word count	$\alpha = \frac{n_k}{k}$	$f(1)$		$f(2)$		$f(3)$	
		Actual	Estimate	Actual	Estimate	Actual	Estimate
Ulysses (Hanley, 1937)	0·115	16,432	15,850	4,776	4,870	2,194	2,220
Eldridge (Good, 1953)	0·136	2,976	3,220	1,079	977	516	400

B. *Scientific publications*

At least four sets of data are available on the number, $f(i)$, of authors contributing a given number, i, of papers each to a journal or journals (Davis, 1941; Leavens, 1953). These are counts of (*a*) papers written by members of the Chicago Section of the American Mathematical Society over a 25-year period; (*b*) papers listed in *Chemical Abstracts* (under A and B) over 10 years; (*c*) papers referred to in a history of physics; and (*d*) papers and abstracts in *Econometrica* over a 20-year period.

We may postulate a mechanism like that of § III, equation (3·1). The authorship of the next paper to appear is 'selected' by stratified sampling from the strata of authors who have previously published 1, 2, ..., papers, the probability for each stratum being proportional to $if(i)$. Again, the probabilities for individual authors need not be proportional to i, but only the probabilities for the aggregates of authors with the same i. For example (as in the case of words), the probability for a particular author may be higher if he has published recently than if he has not. The gradual retirement of authors corresponds to assumption (III).

A comparison of the actual frequencies, for i from 1 to 10, with the estimated frequencies, derived from (2·11) and (2·21), is shown in Table 3. The fit is reasonably good, when it is remembered that only one parameter is available for adjustment. However, it should be

noted that the estimated frequencies tend to be too high for $i = 1, 2$ and too low for $i = 3, \ldots, 10$. In three of the four cases, they are again too high for the tails of the distributions. A further refinement of the model is apparently needed to remove these discrepancies.

Table 3. *Number of persons contributing*

No. of contributions	Chicago Math. Soc.		*Chem. Abstracts*		Physicists		*Econometrica*	
	Actual†	Estimate	Actual†	Estimate	Actual†	Estimate	Actual‡	Estimate
1	133	—	3,991	4,050	784	824	436	453
2	43	46	1,059	1,160	204	217	107	119
3	24	23	493	522	127	94	61	51
4	12	14	287	288	50	50	40	27
5	11	10	184	179	33	30	14	16
6	14	7	131	120	28	20	23	11
7	5	5	113	86	19	14	6	7
8	3	4	85	64	19	10	11	5
9	9 } 13	3 } 10	64	49	6 } 32	8 } 24	1 } 12	4 } 12
10	1	3	65	38	7	6	0	3
11 or more	23	30	419	335	48	52	22	25
Estimated α	0§		0·30		0·39		0·41	
Estimated ρ	0·916§		1·43		1·64		1·69	
k	1,124		22,939		3,396		1,759	
n_k	278		6,891		1,325		721	
ρ'	1·07		—		—		—	

† Davis (1941).
‡ Leavens (1953).
§ $\rho = \bar{\rho}$ estimated in this case from (2·31) to (2·32).

C. *City sizes*

It has been observed, for every U.S. Census since the early nineteenth century, and for most other Western countries as well, that if $F(i)$ is the number of cities of population greater than i, then

$$F(i) \sim Ai^{-\rho}, \tag{4·3}$$

where ρ is close to 1 (see Zipf, 1949, chs. 9, 10).

Again, we would expect such a distribution if the underlying mechanism were one describable by equations like (2·3) and (2·4). Such a mechanism is not hard to conceive. First, equation (2·3) would hold if the growth of population were due solely to the net excess of births over deaths, and if this net growth were proportional to present population. This assumption is certainly satisfied at least roughly. Moreover, it need not hold for each city, but only for the aggregate of cities in each population band. Finally, the equation would still be satisfied if there were net migration to or from cities of particular regions, provided the net addition or loss of population of individual cities *within any region* was proportional to city size. That is, even if all California cities were growing, and all New England cities declining, the equation would hold provided the percentage growth or decline in each area were uncorrelated with city size.

In the case of cities, equation (4·3) could only be expected to hold down to some minimum city size—say, 5000 or 10,000. The constant α would then be interpreted as the fraction of the total population growth in cities above the minimum size that is accounted for by the new cities that reach that size.

D. *Income distribution*

Vilfredo Pareto is generally credited with the discovery that if personal incomes are ranked by size, the number of persons, $F(i)$, whose incomes exceed i can be approximated closely, for the upper ranges of income, by equation (4·3) with ρ usually in the neighbourhood of 1·5 (Davis, 1941; Champernowne, 1953). Hence, the income distributions bear a family resemblance in their upper ranges to those we have already considered, although the parameter, ρ, is substantially larger than 1—its characteristic value in the case of word frequencies and city size distributions.

A stochastic mechanism similar to those described in § III would again produce steady-state distributions closely resembling the observed ones. We picture the stream of income as a sequence of dollars allocated probabilistically to the recipients. If the total annual income of all persons above some specified minimum income is k dollars, the segment of this sequence running from the mth to the $(m+k)$th dollar is the income for the year beginning at time m. We assume that the probability that the next dollar will be allotted to some person with an annual income of i dollars is proportional to $(i+c)f(i)$, with c positive but small. This represents a modification of assumption (I) that decreases the proportion of the total stream going to persons of high income relative to the proportion going to persons with incomes close to the minimum. We assume that a fraction of the dollars is assigned to new persons—i.e. persons reaching the minimum income to which the assumptions apply (assumption (II)). We assume that there is considerable variance among persons within each income class in the probability of receiving additional income, so that the rate at which dollars are dropped from any income class as m increases satisfies assumption (III). Then we obtain again equation (3·8), which now holds for i greater than the minimum income. For large i, this distribution has the required properties with $1/\lambda = \rho$.

The same result has been reached by D. G. Champernowne (1953), following a somewhat different route. He divides income recipients at time t_1 into classes of equal proportionate width. That is, if i_m is the minimum income considered, then the first class contains persons with incomes between i_m and ri_m, the second class, persons with incomes between ri_m and $r^2 i_m$, and so on. Next he introduces transition probabilities p_{gh}, that a person who is in class g at time t_1 will be in class h at time t_2. He assumes that p_{gh} is a function only of $(g-h)$. Now, by his definition of the income classes, the average income of persons in class g will be about $r^{(g-h)}$ times the average income of persons in class h. Hence, the expected income at t_2 of a person who was in class g at t_1 will be

$$\sum_h p_{gh} i_h = \sum_h p_{(g-h)} r^{(g-h)} i_g = \alpha i_g \quad (\alpha \text{ a constant}), \tag{4·4}$$

where i_g is the average income in class g. Prof. Champernowne assumes explicitly that $\alpha < 1$. From this it is clear that his model satisfies our assumptions (I) (in its original form) and (II). Further, since he assumes a substantial variance in income expectations among persons in a given class, our assumption (III) is also approximately satisfied. Hence, in spite of the surface differences between his model and those developed here, the underlying structure is the same.

E. *Biological species*

We conclude this very incomplete list of phenomena exhibiting the Yule distribution by mentioning the example originally analysed by Yule himself (1924). It was discovered by Willis that the number, $f(i)$, of genera of plants having i species each was distributed approximately according to (4·3), with $\rho < 1$. Yule explained these data by a probability model in which the probability, s, of a specific mutation occurring in a particular genus during a short time interval was proportional to the number of species in the genus; while the probability, r, of a generic mutation during the same interval was proportional to the number of genera. Starting at t_0 with a single genus of one species, he computed the distribution $f(i,t)$ for $t_1, t_2, \ldots$, and found the limit as $t \to \infty$. This limiting distribution corresponds to (2·13) with $\rho = r/s$. Yule observed that for $r < s$ (as required to fit the empirical data), this was not a proper distribution function, and obtained the approximate distribution for $t = T$. His procedure was equivalent to replacing the complete Beta function in (2·13) by the incomplete Beta function, taking as the upper limit of integration an appropriate function of T.

If, in the process of § II, we define k as the total number of different species and $f(i,k)$ as the number of genera with exactly i species, we see that our k is a monotonic increasing function of Yule's t (specifically, $k = e^{st}$). Making the appropriate transformation of variables, we find that Yule's assumption with respect to the rate of specific mutation corresponds to our assumption (I′) (and hence is considerably stronger than the assumption we employed in § II). Making the same transformation of variables with respect to his assumption of a constant rate of generic mutation, we find that $n_k = e^{rt}$. We can then compute $\alpha(k)$ (which will now vary with k) by taking the derivative of n_k with respect to k. We obtain

$$\alpha(k) = r\, e^{(r-s)t/s}. \tag{4·5}$$

If we substitute these values in equation (2·34) of case II, where we assumed slowly changing α, we find in the limit, as $t \to \infty$, $\rho = r/s$, as required. Hence, we see that the process of § II is essentially the same as the one treated by Yule.

It is interesting and a little surprising that when Yule, some twenty years after this discovery, examined the statistics of vocabulary, he did not employ this model to account for the observed distributions of word frequencies. Indeed, in his fascinating book on *The Statistical Study of Literary Vocabulary* (1944) he nowhere refers to his earlier paper on biological distributions.

V. CONCLUSION

This paper discusses a number of related stochastic processes that lead to a class of highly skewed distributions (the Yule distribution) possessing characteristic properties that distinguish them from such well-known functions as the negative binomial and Fisher's logarithmic series. In § I, the distinctive properties of the Yule distribution were described. In §§ II and III several stochastic processes were examined from which this distribution can be derived. In § IV, a number of empirical distributions that can be approximated closely by the Yule distribution were discussed, and mechanisms postulated to explain why they are determined by this particular kind of stochastic process. In the same section, the derivations of §§ II and III were compared with models previously proposed by Yule (1924) and Champernowne (1953) to account for the data on biological species and on incomes, respectively.

The probability assumptions we need for the derivations are relatively weak, and of the same order of generality as those commonly employed in deriving other distribution functions—the normal, Poisson, geometric and negative binomial. Hence, the frequency with which the Yule distribution occurs in nature—particularly in social phenomena—should occasion no great surprise. This does not imply that all occurrences of this empirical distribution are to be explained by the process discussed here. To the extent that other mechanisms can be shown also to lead to the same distribution, its common occurrence is the less surprising. Conversely, the mere fact that particular data conform to the Yule distribution and can be given a plausible interpretation in terms of the stochastic model proposed here tells little about the underlying phenomena beyond what is contained in assumptions (I) through (III).

REFERENCES

Champernowne, D. G. (1953). A model of income distribution. *Econ. J.* **63**, 318.

Darwin, J. H. (1953). Population differences between species growing according to simple birth and death processes. *Biometrika*, **40**, 370.

Davis, Harold T. (1941). *The Analysis of Economic Time Series*. Principia Press.

Feller, William (1950). *An Introduction to Probability Theory and Its Applications*, vol. **1**. Wiley.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237.

Hanley, Miles L. (1937). *Word Index to James Joyce's Ulysses*. University of Wisconsin Press.

Kendall, David G. (1948). On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika*, **35**, 6.

Leavens, Dickson H. (1953). *Econometrica*, **21**, 630.

Mandelbrot, Benoit (1953). An informational theory of the statistical structure of language. In *Communication Theory* (ed. by Willis Jackson). Butterworths.

Thorndike, Edward L. (1937). On the number of words of any given frequency of use. *Psychol. Rec.* **1**, 399.

Titchmarsh, E. C. (1939). *The Theory of Functions*, 2nd ed. Oxford University Press.

Yule, G. Udny (1924). A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S. *Phil. Trans.* B, **213**, 21.

Yule, G. Udny (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.

SIMULTANEOUS TESTS OF LINEAR HYPOTHESES

By M. N. GHOSH

University of North Carolina and University of Calcutta

1. Introduction

A very common situation in the analysis of variance of survey data, where the investigator is not able to put his experimental data in the framework of a design, planned in advance, is that the estimates of the parameters are correlated in various ways and the analysis of variance becomes cumbersome. Even for the analysis of variance of a two-way classification with unequal class frequencies one meets with this difficulty. We shall be concerned here with the analysis of such data and the test of significance for groups of parameters representing different aspects of the problem. We consider below a concrete example, in which standing height, sexual maturity characters and certain blood chemicals, like haemoglobin, ascorbic acid, carotene and alkaline phosphates were measured for a number of girls between the ages 9 and 14 years.* For the purpose of analysis the girls were divided into twenty-five classes according to sexual maturity ratings from indices of breast (B_i) and pubic hair development (Ph_j), each of these being a 5-point rating.

The following model is assumed:

$$y_{k(i,j)} = \mu_{ij} + \beta_1 x_{1,k} + \beta_2 x_{2,k} + \beta_3 x_{3,k} + \gamma z_k + \epsilon_k, \tag{1·1}$$

where $(y_k, x_{1,k}, x_{2,k}, x_{3,k}, z_k)$ $(k = 1, 2, \ldots, N)$ are sample individuals. $y_{k(i,j)}$ represents the height increment of the kth girl in the maturity class $(B_i Ph_j)$, $x_{1,k}, x_{2,k}, x_{3,k}$ are nutritional variables haemoglobin, ascorbic acid and carotene respectively, and z_k is alkaline phosphatase. We thus consider three groups of parameters $\{\mu_{ij}\}$ $(i,j = 1, \ldots, 5)$, $\{\beta_1, \beta_2, \beta_3\}$ and $\{\gamma\}$. The hypotheses corresponding to these three groups of parameters are

$$\left.\begin{aligned} H_1&: \ \mu_{ij} = \mu \quad (i,j = 1, 2, \ldots, 5), \\ H_2&: \ \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \\ H_3&: \ \gamma = 0. \end{aligned}\right\} \tag{1·2}$$

Rejection of the hypothesis H_1 would imply that the height increment depends upon sexual maturity, the rejection of H_2 would imply that the blood chemicals, which depend upon nutritional status, affect height increment, and the rejection of H_3 would imply that alkaline phosphatase affects height increment.

The hypotheses H_1, H_2, H_3 may be tested by the analysis of variance method, but these tests are not independent, because of the non-orthogonality of the groups of estimates as well as for the fact that we have to use the same estimate of error variance. We shall introduce the notion of quasi-independent tests of multiple hypotheses, which is appropriate in such situations.

* The data were collected by Dr Hughes Bryan, Professor of Nutrition, School of Public Health, University of North Carolina.

2. Quasi-independent tests of multiple hypotheses

For any test of significance we consider the first and second kinds of error. Let the hypotheses tested be

$$H_1\colon\ \theta_1 = 0, \quad H_2\colon\ \theta_2 = 0, \quad H_3\colon\ \theta_3 = 0. \tag{2·1}$$

The tests T_1, T_2, T_3 of H_1, H_2, H_3 will be called quasi-independent when

$$\left.\begin{aligned} \text{Prob}_{T_1}\{\text{acceptance } H_1 \mid \theta_1 \neq 0, \theta_2, \theta_3\} &= \text{Prob}_{T_1}\{\text{acceptance } H_1 \mid \theta_1 \neq 0, \theta_2 = 0, \theta_3 = 0\}, \\ \text{Prob}_{T_1}\{\text{rejection } H_1 \mid \theta_1 = 0, \theta_2, \theta_3\} &= \text{Prob}_{T_1}\{\text{rejection } H_1 \mid \theta_1 = 0, \theta_2 = 0, \theta_3 = 0\}, \end{aligned}\right\} \tag{2·2}$$

hold for all values of θ_2 and θ_3, i.e. whether H_2 and H_3 are true or not; and similar equations hold for T_2 and T_3. where $\text{Prob}_T\ \{\ \}$ denotes the probability of the statement in parentheses for the test procedure T_1, etc. Thus for quasi-independent tests of hypotheses H_1, H_2 and H_3 the first and second kinds of error for each hypothesis do not depend on the parameters of the other hypotheses.

Two tests of hypotheses H_1 and H_2 with the critical regions C_1 and C_2 are independent if the following holds for all θ_1 and θ_2:

$$\text{Prob}\{X \subset C_1,\ X \subset C_2 \mid \theta_1, \theta_2\} = \text{Prob}\{X \subset C_1 \mid \theta_1\},\ \text{Prob}\{X \subset C_2 \mid \theta_2\},$$

X being the observed sample.

It will be seen later that the usual analysis of variance tests for multiple hypotheses are quasi-independent, and we most often do not need independent tests.

3. Control of errors in the simultaneous tests of hypotheses

There may be different points of view for assigning significance levels, in the case of simultaneous tests of hypotheses. In certain situations, where the decisions regarding the hypotheses $H_1, H_2, \ldots$, etc., are unrelated, it is proper to consider the significance level of each hypothesis individually at 5 or 1 % level (say). But when the decisions have a joint import, it is proper to consider the first kind of error of the simultaneous test of H_1 and H_2, as the rejection of at least one of the hypotheses, when all of them are in fact true.

The significance level of the simultaneous test is defined as the probability of rejecting at least one of the hypotheses, when all of them are true.

Suppose in the test of a new variety of crop against a common variety used as a control we are interested in two different characters and the new variety may be considered desirable when it is superior in either of these characters. Since one would be interested to know which particular character in the new variety is superior to the control, a simultaneous test should be done. In this case, the significance level, as some sort of measure of the amount of caution implied in the test, should naturally be that of the simultaneous test, since one would like to insure, at a certain level, against declaring the new variety superior, when it is actually no better than the control.

A similar situation exists in quality control for acceptance of material, when a number of characters are examined and the material is rejected when it is not up to the mark with respect to any of these characters. It may be useful, in this case, to determine the particular deficiency in any of these characters, and so we must use a simultaneous test, and the significance level of the simultaneous test should be used to insure against too frequent rejections. Of course, we pay for this by widening confidence intervals for the parameters measuring these characters.

The need for a safety device like the simultaneous significance level would be even more apparent for the agricultural or quality control example stated above, when the number of characters is large, in which case the chance of declaring a new variety different from the common variety may be much larger than 5 %. However, the control of the first kind of error in a simultaneous test is achieved only at the expense of increasing the second kind of error for individual characters. In any particular problem, whether the point of view of individual or simultaneous significance level should be adopted depends, roughly, on how much *a priori* weight one attaches to the alternative hypotheses, either from theoretical expectations or from considerations of cost in replacing the old variety by the new variety. If this cost is small, e.g. if the varieties are grown only on an experimental scale, one would be relatively free to decide on either of the varieties as superior and the individual significance level would be the proper one. However, with varieties established in agricultural practice, any statistical decision in favour of a new variety would involve large expenditures and the statistician should take an attitude of caution. The simultaneous significance level takes account of this attitude and is thus appropriate in such cases.

In the analysis of variance problem considered before, the conventional statistical procedure of (*a*) testing for the overall SS for fitting constants, (*b*) testing separately for the SS of H_1, adjusting for H_2 at 5 % level, etc., would give quasi-independent tests at individual levels of significance 5 %. But the problem of obtaining the simultaneous significance level in this case is mathematically a very intractable one, and we shall find an upper bound for the significance level. This would give a control of the first kind of error. The operating characteristic would, of course, be of the same nature as for the usual analysis of variance of multiple hypotheses, since the test procedure is essentially the same, except that the exact significance level is not known but only its upper bound.

The notion of simultaneous level of significance has already been considered in various ways and languages by Scheffé (1953), Tukey (1952) and Nandi (1951), and the practical implications have been thoroughly discussed by Tukey in a mimeographed report.

4. Simultaneous tests and significance levels

Consider the linear model

$$E(y_i) = a_{i1}p_1 + \ldots + a_{im}p_m \quad (i = 1, 2, \ldots, N), \tag{4·1}$$

y_i being independent normal variables with unknown variance σ^2, and $p_1, \ldots, p_m$ are unknown parameters. Let rank $(a_{ij}) = N_0$ and let $\pi_1 \ldots \pi_R$ be estimable linear functions of the parameters $p_1, \ldots, p_m$

$$\pi_\kappa = l_{\kappa 1}p_1 + l_{\kappa 2}p_2 + \ldots + l_{\kappa m}p_m \quad (\kappa = 1, 2, \ldots, R), \tag{4·2}$$

such that the coefficient vectors $(l_{\kappa 1}, \ldots, l_{\kappa m})$ form a vector space of rank $R \leqslant N_0$. We consider the following linear hypotheses,

$$\left.\begin{aligned} H_1&: \quad \pi_1 = 0, \ \ldots, \ \pi_{\kappa_1} = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ H_s&: \quad \pi_{\kappa_1+\ldots+\kappa_{s-1}+1} = 0, \ \ldots, \ \pi_{\kappa_1+\ldots+\kappa_s} = 0. \end{aligned}\right\} \tag{4·3}$$

Let $Y_1, \ldots, Y_{\kappa_1}; Y_{\kappa_1+1}, \ldots, Y_{\kappa_1+\kappa_2}, \ldots, Y_{\kappa_1+\ldots+\kappa_s}$ be the best linear estimates of these parameters, obtained by the method of least squares. We shall sometimes denote the coefficient vectors

of these linear functions by the same symbol, so that we have the alternative notation for the linear function $Y_i = (Y_i, y)$, where (Y_i, y) is the scalar product of the vector Y_i and the vector y. From Markoff's theorem we have an independent estimate of error variance, S_e^2, with n_e d.f., which is independent of the parameters $p_1, \ldots, p_m$. Quasi-independent tests of the hypotheses H_n can be made by considering the linear functions

$$(Y_{b_{n-1}+1}, y) \ldots (Y_{b_n}, y) \quad (b_n = \kappa_1 + \ldots + \kappa_n),$$

whose expectations are $\pi_{b_{n-1}+1}, \ldots, \pi_{b_n}$. Let $U_{b_{n-1}+1}, \ldots, U_{b_n}$ be orthonormal vectors forming a basis of the vector space formed by $Y_{b_{n-1}+1}, \ldots, Y_{b_n}$. Then (U_i, y) is a linear form in (Y_j, y) $(i, j = b_{n-1}+1, \ldots, b_n)$ and

$$E\{(U_i, y)\} = \sum_{j=b_{n-1}+1}^{b_n} \alpha_j E(Y_j, y) = \sum_{j=b_{n-1}+1}^{b_n} \alpha_j \pi_j \quad \text{from} \quad E(Y_j, y) = \pi_j. \tag{4·4}$$

On the hypothesis H_n, $\chi_n^2/\sigma^2 = \sum_{j=b_{n-1}+1}^{b_n} [(U_j, y) - E(U_j, y)]^2$ has a χ^2-distribution with κ_n d.f. and $\chi_n^2/(\kappa_n S_e^2)$ has an F-distribution with d.f. (r, n_e). The second kind of error for the test of H_n depends only upon the parameters $\pi_{b_{n-1}+1}, \ldots, \pi_{b_n}$.

We now consider sets $M_1, M_2, \ldots$ of vectors $Y_1, \ldots, Y_{\kappa_1}; Y_{\kappa_1+1}, \ldots, Y_{\kappa_1+\kappa_2}; \ldots$ belonging to hypotheses $H_1, \ldots, H_s$, so that all vectors belonging to the same hypothesis H_i belong to the same set M_i. Vectors belonging to different sets are orthogonal, and if the vectors belonging to two hypotheses H_i and H_j are orthogonal they belong to different sets. These sets we shall call orthogonal sets and hypotheses belonging to two different sets, orthogonal hypotheses. We shall consider different cases according as an orthogonal set consists of a single hypothesis H_i or more.

Case I. Let all hypotheses $H_1, \ldots, H_s$ be orthogonal so that each orthogonal set M_i consists of a single hypothesis. Let $U_1, \ldots, U_{\kappa_1}; U_{\kappa_1+1}, \ldots, U_{\kappa_1+\kappa_2}; \ldots, U_{\kappa_1+\ldots+\kappa_s}$ be orthonormal systems of vectors in the spaces determined by $Y_1, \ldots, Y_{\kappa_1}; Y_{\kappa_1+1}, \ldots, Y_{\kappa_1+\kappa_2}; \ldots, Y_{\kappa_1+\ldots+\kappa_s}$ respectively. Let

$$E\{(U_i, y)\} = \Phi_i.$$

We shall consider for simplicity $s = 3$. The joint distribution of

$$\chi_1^2 = \sum_{i=1}^{\kappa_1} \{(U_i, y) - \Phi_i\}^2, \quad \chi_2^2 = \sum_{i=\kappa_1+1}^{\kappa_1+\kappa_2} \{(U_i, y) - \Phi_i\}^2, \quad \chi_3^2 = \sum_{i=\kappa_1+\kappa_2+1}^{\kappa_1+\kappa_2+\kappa_3} \{(U_i, y) - \Phi_i\}^2, \tag{4·5}$$

and S_e^2 is given by

$$\text{const.} \exp\left\{-\frac{1}{2\sigma^2}[\chi_1^2 + \chi_2^2 + \chi_3^2 + n_e S_e^2]\right\} (\chi_1^2)^{\frac{1}{2}(\kappa_1-2)} (\chi_2^2)^{\frac{1}{2}(\kappa_2-2)} (\chi_3^2)^{\frac{1}{2}(\kappa_3-2)} (S_e^2)^{\frac{1}{2}(n_e-2)} d\chi_1^2 d\chi_2^2 d\chi_3^2 dS_e^2.$$

Put

$$\frac{\chi_1^2}{n_e S_e^2} = G_1, \quad \frac{\chi_2^2}{n_e S_e^2} = G_2, \quad \frac{\chi_3^2}{n_e S_e^2} = G_3.$$

Making the transformation, integration for S_e^2 gives the distribution of G_1, G_2, G_3 as

$$C(\kappa_1, \kappa_2, \kappa_3; n_e) \frac{G_1^{\frac{1}{2}(\kappa_1-2)} G_2^{\frac{1}{2}(\kappa_2-2)} G_3^{\frac{1}{2}(\kappa_3-2)}}{(1+G_1+G_2+G_3)^{\frac{1}{2}(\kappa_1+\kappa_2+\kappa_3+n_e)}} dG_1 dG_2 dG_3. \tag{4·6}$$

We now consider the region defined by $G_1 < \lambda_1$, $G_2 < \lambda_2$, $G_3 < \lambda_3$ which has the probability $P(\lambda_1, \lambda_2, \lambda_3)$, say. If $P(\lambda_1, \lambda_2, \lambda_3) = 1-\alpha$, we have the system of simultaneous confidence regions $C_1, C_2, \ldots$ for the sets of parameters with the confidence coefficient $1-\alpha$, as

$$\left.\begin{aligned} C_1: \quad & \pi_1, \ldots, \pi_{\kappa_1}: \quad \sum_{i=1}^{\kappa_1} \{(U_i, y) - \Phi_i\}^2 < \lambda_1 n_e S_e^2, \\ C_2: \quad & \pi_{\kappa_1+1}, \ldots, \pi_{\kappa_1+\kappa_2}: \quad \sum_{i=\kappa_1+1}^{\kappa_1+\kappa_2} \{(U_i, y) - \Phi_i\}^2 < \lambda_2 n_e S_e^2, \\ C_3: \quad & \pi_{\kappa_1+\kappa_2+1}, \ldots, \pi_{\kappa_1+\kappa_2+\kappa_3}: \quad \sum_{i=\kappa_1+\kappa_2+1}^{\kappa_1+\kappa_2+\kappa_3} \{(U_i, y) - \Phi_i\}^2 < \lambda_3 n_e S_e^2. \end{aligned}\right\} \tag{4·7}$$

The best choice of $\lambda_1, \lambda_2, \lambda_3$, i.e. to give the smallest confidence regions, is not known and needs further investigation. From considerations of degrees of freedom we may consider $\lambda_1/\kappa_1 = \lambda_2/\kappa_2 = \lambda_3/\kappa_3 = \lambda$.

5. Evaluation of significance level

To determine λ from (4·6) we have

$$C(\kappa_1, \kappa_2, \kappa_3; n_e) \int_0^{\kappa_1\lambda} dG_1 \int_0^{\kappa_2\lambda} dG_2 \int_0^{\kappa_3\lambda} dG_3 \frac{G_1^{\frac{1}{2}(\kappa_1-2)} G_2^{\frac{1}{2}(\kappa_2-2)} G_3^{\frac{1}{2}(\kappa_3-2)}}{(1+G_1+G_2+G_3)^{\frac{1}{2}(\kappa_1+\kappa_2+\kappa_3+n_e)}} = 1-\alpha, \tag{5·1}$$

where α is the significance level. Put $t_1 = G_1$, $t_2 = G_2/(1+G_1)$, $t_3 = G_3/(1+G_1+G_2)$, then the above becomes

$$C(\kappa_1, \kappa_2, \kappa_3; n_e) \int_0^{\kappa_1\lambda} \frac{t_1^{\frac{1}{2}(\kappa_1-2)} dt_1}{(1+t_1)^{\frac{1}{2}(\kappa_1+n_e)}} \left[\int_0^{\kappa_2\lambda/(1+t_1)} \frac{t_2^{\frac{1}{2}(\kappa_2-2)} dt_2}{(1+t_2)^{\frac{1}{2}(\kappa_1+\kappa_2+n_e)}} \int_0^{\kappa_3\lambda/[(1+t_1)(1+t_2)]} \frac{t_3^{\frac{1}{2}(\kappa_3-2)} dt_3}{(1+t_3)^{\frac{1}{2}(\kappa_1+\kappa_2+\kappa_3+n_e)} dt_3} \right]. \tag{5·2}$$

which can be evaluated by successive numerical integration and using Pearson's Tables of Incomplete B-functions.

For any confidence coefficient $1-\alpha$, the calculation of λ depends upon calculations of iterated integrals, and tables have to be prepared for these. One difficulty of preparing such tables is that the integrals depend upon the parameters $\kappa_1, \kappa_2, \kappa_3; n_e$, etc., and thus unless the number of parameters can be reduced, construction of tables would be difficult. We may, however, get an upper bound for the significance level α, from an inequality of Kimball (1951) given below:

$$\Pr\{|G_1| \leqslant \lambda_1, |G_2| \leqslant \lambda_2, |G_3| \leqslant \lambda_3\} \geqslant \Pr\{|G_1| \leqslant \lambda_1\} \Pr\{|G_2| \leqslant \lambda_2\} \Pr\{|G_3| \leqslant \lambda_3\}. \tag{5·3}$$

Here we choose $\lambda_1, \lambda_2, \lambda_3$ so that

$$\Pr\{|G_1| \leqslant \lambda_1\} = \Pr\{|G_2| \leqslant \lambda_2\} = \Pr\{|G_3| \leqslant \lambda_3\}. \tag{5·4}$$

Thus if $\Pr\{|G_1| \leqslant \lambda_1, |G_2| \leqslant \lambda_2, |G_3| \leqslant \lambda_3\} = 0{\cdot}95$ we have to make $\Pr\{|G_1| \leqslant \lambda_1\} = 0{\cdot}983$. Now $n_e G_1$ has an F-distribution with κ_1 and n_e d.f. so that $\lambda_1 = \frac{\kappa_1}{n_e} F_\alpha(\kappa_1, n_e)$, where $F_\alpha(\kappa, n_e)$ is the α% point of the F-distribution with d.f. (κ_1, n_e). We thus have

$$\Pr\left\{|G_1| \leqslant \frac{\kappa_1}{n_e} F_{\alpha'}(\kappa_1, n_e), |G_2| \leqslant \frac{\kappa_2}{n_e} F_{\alpha'}(\kappa_2, n_e), |G_3| \leqslant \frac{\kappa_3}{n_e} F_{\alpha'}(\kappa_3, n_e)\right\} \geqslant (1-\alpha')^3 = 1-\alpha \quad \text{say.}$$

Thus

$$\alpha' = 1-(1-\alpha)^{\frac{1}{3}}. \tag{5·5}$$

Case II. We now consider a set M_i, which consists of more than one group of linear functions, say it contains linear functions corresponding to the hypotheses H_1 and H_2. We shall call this a compound set. In this case the linear functions belonging to H_1 and H_2 are non-orthogonal. We shall show that there is no non-singular linear transformation by which these linear functions can be transformed into mutually orthogonal sets corresponding to hypotheses H_1 and H_2 respectively. Thus the basic inequality (5·3) is not directly applicable.

Let $U_1, \ldots, U_{\kappa_1}; U_{\kappa_1+1}, \ldots, U_{\kappa_1+\kappa_2}$ be an orthonormal basis of the vector space formed by $Y_1, \ldots, Y_{\kappa_1}; Y_{\kappa_1+1}, \ldots, Y_{\kappa_1+\kappa_2}$, so that $U_1, \ldots, U_{\kappa_1}$ is a basis of the vector space formed by $Y_1, \ldots, Y_{\kappa_1}$. As before, let $E\{(U_i, y)\} = \Phi_i$, then Φ_i are linear functions of $\pi_1, \ldots, \pi_{\kappa_1}; \pi_{\kappa_1+1}, \ldots, \pi_{\kappa_1+\kappa_2}$. Let $\bar{Y}_i$ be the normalized vector corresponding to Y_i, i.e. $\bar{Y}_i = \dfrac{Y_i}{|Y_i|}$, where $|Y_i|$ is the norm of the vector Y_i and let $\bar{\pi}_i = \dfrac{\pi_i}{|Y_i|}$. The relation between U-vectors and Y-vectors is expressed by

$$\begin{pmatrix} \bar{Y}_{\mathrm{I}} \\ \bar{Y}_{\mathrm{II}} \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ \beta & \gamma \end{pmatrix} \begin{pmatrix} U_{\mathrm{I}} \\ U_{\mathrm{II}} \end{pmatrix}, \tag{5·6}$$

where $\alpha = (\alpha_{i,j})$ is a $(\kappa_1 \times \kappa_1)$ matrix, $\gamma = (\gamma_{ij})$ is a $(\kappa_2 \times \kappa_2)$ matrix, $\beta = (\beta_{ij})$ is a $(\kappa_2 \times \kappa_1)$ matrix, $\bar{Y}_{\mathrm{I}}$ is a $(1 \times \kappa_1)$ matrix with components $\bar{Y}_1, \ldots, \bar{Y}_{\kappa_1}$, $\bar{Y}_{\mathrm{II}}$ is a $(1 \times \kappa_2)$ matrix with components $\bar{Y}_{\kappa_1+1}, \ldots, \bar{Y}_{\kappa_1+\kappa_2}$ and similarly U_{I} is a $(1 \times \kappa_1)$ matrix with components $U_1, \ldots, U_{\kappa_1}$, etc. From the choice of the basis $U_1, \ldots, U_{\kappa_1}; U_{\kappa_1+1}, \ldots, U_{\kappa_1+\kappa_2}$, it is clear that the matrices α and γ are non-singular and $\beta \neq 0$, since the vectors $Y_1, \ldots, Y_{\kappa_1}$ and $Y_{\kappa_1+1}, \ldots, Y_{\kappa_1+\kappa_2}$ are non-orthogonal. Inverting the equation (5·6) we get

$$\begin{pmatrix} U_{\mathrm{I}} \\ U_{\mathrm{II}} \end{pmatrix} = \begin{pmatrix} \alpha^{-1}, & 0 \\ \gamma^{-1}\beta\alpha^{-1}, & \gamma^{-1} \end{pmatrix} \begin{pmatrix} \bar{Y}_{\mathrm{I}} \\ \bar{Y}_{\mathrm{II}} \end{pmatrix}. \tag{5·7}$$

Replacing the vectors $\bar{Y}_i$ by $\bar{\pi}_i$ and U_i by Φ_i, the corresponding relation must hold between the Φ's and the $\bar{\pi}$'s. Hence $\Phi_1, \ldots, \Phi_{\kappa_1}$ depend upon $\bar{\pi}_1, \ldots, \bar{\pi}_{\kappa_1}$ alone but $\Phi_{\kappa_1+1}, \ldots, \Phi_{\kappa_1+\kappa_2}$ cannot be expressed solely in terms of $\bar{\pi}_{\kappa_1+1}, \ldots, \bar{\pi}_{\kappa_1+\kappa_2}$, since $(\gamma^{-1}\beta\alpha^{-1}) \neq 0$. Thus the sets of linear functions $U_1, \ldots, U_{\kappa_1}$ and $U_{\kappa_1+1}, \ldots, U_{\kappa_1+\kappa_2}$ do not provide quasi-independent tests of H_1 and H_2. We shall see later in II (*c*) that this can be done when the transformation is singular, i.e. when we chose a suitable subset of $\kappa_2 - \kappa_1$ vectors from the vectors $U_{\kappa_1+1}, \ldots, U_{\kappa_1+\kappa_2}$.

We shall also consider two other methods of setting inequalities to reduce the problem to orthogonal sets. The essence of these methods is to cover a complicated region C by a sphere S and using the probability of the sphere for the region C as an upper bound.

Method (a). Method of simultaneous confidence intervals for all contrasts

This method is due to Scheffe (1953), Tukey (1952) and Roy & Bose (1953). It can be used for the simultaneous test but it also gives the confidence intervals for all linear functions of the parameters $\pi_1, \ldots, \pi_{\kappa_1}; \pi_{\kappa_1+1}, \ldots, \pi_{\kappa_1+\kappa_2}$ simultaneously with a given confidence coefficient. In equation (5·6) since both U's and $\bar{Y}$'s are normalized vectors

$$\left.\begin{aligned} \sum_{i=1}^{\kappa_1} \alpha_{ni}^2 &= 1 \quad (n = 1, \ldots, \kappa_1), \\ \sum_{i=1}^{\kappa_1} \beta_{ni}^2 + \sum_{i=1}^{\kappa_2} \gamma_{ni}^2 &= 1 \quad (n = 1, \ldots, \kappa_2). \end{aligned}\right\} \tag{5·8}$$

Also

$$\left.\begin{aligned} (\bar{Y}_n, y) &= \sum_{i=1}^{\kappa_1} \alpha_{ni}(U_i, y) && (n = 1, 2, \ldots, \kappa_1), \\ (\bar{Y}_{\kappa_1+n}, y) &= \sum_{i=1}^{\kappa_1} \beta_{ni}(U_i, y) + \sum_{i=1}^{\kappa_2} \gamma_{ni}(U_{\kappa_1+n}, y) && (n = 1, 2, \ldots, \kappa_2), \end{aligned}\right\} \tag{5·9}$$

and similar equations hold between π's and Φ's. From Schwartz's inequality

$$[(\tilde{Y}_n, y) - \tilde{\pi}_n]^2 \leqslant \sum_{j=1}^{\kappa_1+\kappa_2} [(U_j, y) - \Phi_j]^2 \quad (n = 1, 2, \ldots, \kappa_1+\kappa_2).$$

Hence
$$\Pr\{|(\tilde{Y}_n; y) - \tilde{\pi}_n| \leqslant \delta \ (n = 1, \ldots, \kappa_1+\kappa_2)\} \geqslant \Pr\left\{\sum_{j=1}^{\kappa_1+\kappa_2} [(U_j, y) - \Phi_j]^2 \leqslant \delta^2\right\}. \tag{5·10}$$

For the set M_t we now use $\chi_t^2 = \sum_{j=1}^{\kappa_1+\kappa_2} [(U_j, y) - \Phi_j]^2$ in (4·7). This method has the advantage that if π is a linear function of $\tilde{\pi}_i$'s, then at the same time, with the same confidence coefficient, we have the confidence interval for π,

$$\left.\begin{aligned} \pi &= \alpha_1\tilde{\pi}_1 + \ldots + \alpha_{\kappa_1+\kappa_2}\tilde{\pi}_{\kappa_1+\kappa_2}, \\ \alpha_1\tilde{Y}_1 + \ldots + \alpha_{\kappa_1+\kappa_2}\tilde{Y}_{\kappa_1+\kappa_2} &\pm \delta\sqrt{(\Sigma\alpha_i^2)}. \end{aligned}\right\} \tag{5·11}$$

Method (b)

From the equation (5·6) we have
$$\tilde{Y}_{II} = \beta U_I + \gamma U_{II}. \tag{5·12}$$

Since the matrix γ is non-singular

$$\gamma^{-1}\tilde{Y}_{II} = \gamma^{-1}\beta U_I + U_{II}. \tag{5·13}$$

Let $\gamma^{-1}\tilde{Y}_{II} = z_{II}$, which is a column vector with components $z_{\kappa_1+1}, \ldots, z_{\kappa_1+\kappa_2}$, and let $\tilde{z}_{\kappa_1+1}, \ldots, \tilde{z}_{\kappa_1+\kappa_2}$ be the corresponding normalized vectors, then

$$\left.\begin{aligned} (\tilde{z}_{\kappa_1+1}, y) &= \beta'_{1,1}(U_1, y) + \ldots + \beta'_{1,\kappa_1}(U_{\kappa_1}, y) + \gamma'_1(U_{\kappa_1+1}, y), \\ &\vdots \\ (\tilde{z}_{\kappa_1+\kappa_2}, y) &= \beta'_{\kappa_2,1}(U_1, y) + \ldots + \beta'_{\kappa_2,\kappa_1}(U_{\kappa_1}, y) + \gamma'_{\kappa_2}(U_{\kappa_1+\kappa_2}, y), \end{aligned}\right\} \tag{5·14}$$

so that $\sum_{i=1}^{\kappa_1} \beta'^2_{ni} + \gamma'^2_n = 1$ for $n = \kappa_1+1, \ldots, \kappa_1+\kappa_2$, and similar equations hold between the expected values $\psi_j = E(\tilde{z}_j, y)$ and $\Phi_i = E(U_i, y)$. Hence from Schwartz's inequality

$$\left.\begin{aligned} [(\tilde{z}_n, y) - \psi_n]^2 &\leqslant \sum_{i=1}^{\kappa_1} [(U_i, y) - \Phi_i]^2 + [(U_n, y) - \Phi_n]^2 \quad (n = \kappa_1+1, \ldots, \kappa_1+\kappa_2), \\ \sum_{n=\kappa_1+1}^{\kappa_1+\kappa_2} [(\tilde{z}_n, y) - \psi_n]^2 &\leqslant \kappa_2 \sum_{n=1}^{\kappa_1} [(U_i, y) - \Phi_i]^2 + \sum_{n=\kappa_1+1}^{\kappa_1+\kappa_2} [(U_n, y) - \Phi_n]^2. \end{aligned}\right\} \tag{5·15}$$

Since $\{z_n\}$ are linear functions of $\tilde{Y}_{\kappa_1+1}, \ldots, \tilde{Y}_{\kappa_1+\kappa_2}$, $\psi_n = E(\tilde{z}_n, y)$ $(n = \kappa_1+1, \ldots)$ are linear functions of parameters $\pi_{\kappa_1+1}, \ldots, \pi_{\kappa_1+\kappa_2}$, a confidence region for these parameters is given by

$$\sum_{n=\kappa_1+1}^{\kappa_1+\kappa_2} [(\tilde{z}_n, y) - \psi_n]^2 \leqslant \kappa_2 \sum_{n=1}^{\kappa_1} [(U_i, y) - \Phi_i]^2 + \sum_{n=\kappa_1+1}^{\kappa_1+\kappa_2} [(U_n, y) - \Phi_n]^2 \leqslant C_2, \tag{5·16}$$

while a confidence region for the parameters $\pi_1, \ldots, \pi_{\kappa_1}$ is given by

$$\sum_{n=1}^{\kappa_1} [(U_n, y) - \Phi_i]^2 \leqslant C_1. \tag{5·17}$$

When there are more than two hypotheses in the set M_t, e.g. if there are three hypotheses H_1, H_2, H_3 with the groups of parameters $\pi_1, \ldots, \pi_{\kappa_1}$; $\pi_{\kappa_1+1}, \ldots, \pi_{\kappa_1+\kappa_2}$; $\pi_{\kappa_1+\kappa_2+1}, \ldots, \pi_{\kappa_1+\kappa_2+\kappa_3}$ with best linear estimates $Y_1, \ldots, Y_{\kappa_1}$; $Y_{\kappa_1+1}, \ldots, Y_{\kappa_1+\kappa_2}$; $Y_{\kappa_1+\kappa_2+1}, \ldots, Y_{\kappa_1+\kappa_2+\kappa_3}$, we use the same

method as before for the hypothesis H_3 and by combining the hypotheses H_1 and H_2. We then have the following confidence regions for the three groups of parameters:

$$\left.\begin{aligned}
&(\pi_1, \ldots, \pi_{\kappa_1}): \sum_{i=1}^{\kappa_1} [(U_i, y) - \Phi_i]^2 \leqslant C_1, \\
&(\pi_{\kappa_1+1}, \ldots, \pi_{\kappa_1+\kappa_2}): \sum_{n=\kappa_1+1}^{\kappa_1+\kappa_2} [(z_n, y) - \psi_n]^2 \leqslant \kappa_2 \sum_{i=1}^{\kappa_1} [(U_i, y) - \Phi_i]^2 + \sum_{i=\kappa_1+1}^{\kappa_1+\kappa_2} [(U_i, y) - \Phi_i]^2 \leqslant C_2, \\
&(\pi_{\kappa_1+\kappa_2+1}, \ldots, \pi_{\kappa_1+\kappa_2+\kappa_3}): \sum_{n=\kappa_1+\kappa_2+1}^{\kappa_1+\kappa_2+\kappa_3} [(z_n, y) - \psi_n]^2 \leqslant \kappa_3 \sum_{i=1}^{\kappa_1+\kappa_2} [(U_i, y) - \Phi_i]^2 + \sum_{i=\kappa_1+\kappa_2+1}^{\kappa_1+\kappa_2+\kappa_3} [(U_i, y) - \Phi_i]^2 \leqslant C_3.
\end{aligned}\right\} \tag{5·18}$$

For the test of the hypotheses

$$\begin{aligned}
H_1: &\quad \pi_1 = 0, \quad \ldots, \quad \pi_{\kappa_1} = 0 \\
H_2: &\quad \pi_{\kappa_1+1} = 0, \quad \ldots, \quad \pi_{\kappa_1+\kappa_2} = 0 \\
H_3: &\quad \pi_{\kappa_1+\kappa_2+1} = 0 \quad \ldots, \quad \pi_{\kappa_1+\kappa_2+\kappa_3} = 0
\end{aligned}$$

we consider

$$\left.\begin{aligned}
&\chi_1^2 = \sum_{i=1}^{\kappa_1} (U_i, y)^2, \quad F_1 = \chi_1^2/(\kappa_1 S_e^2), \\
&\chi_2^2 = \sum_{i=\kappa_1+1}^{\kappa_1+\kappa_2} (U_i, y)^2, \quad F_2 = \chi_2^2/(\kappa_2 S_e^2). \\
&\chi_3^2 = \sum_{i=\kappa_1+\kappa_2+1}^{\kappa_1+\kappa_2+\kappa_3} (U_i, y)^2, \quad F_3 = \chi_3^2/(\kappa_3 S_e^2),
\end{aligned}\right\} \tag{5·19}$$

Since $\Phi_i = 0$ on the given hypotheses, for all $i = 1, 2, \ldots, \kappa_1 + \kappa_2 + \kappa_3$. If the $\alpha\,\%$ point of F_1, F_2, F_3 (corresponding to G_1, G_2, G_3) are ϵ_1, ϵ_2, ϵ_3, then we consider the significance limits of $\frac{1}{\kappa_1}\chi_1^2, \frac{1}{\kappa_2}\chi_2^2, \frac{1}{\kappa_3}\chi_3^2$ for the test of hypotheses H_1, H_2, H_3 respectively as $n_e S_e^2 \epsilon_1$, $n_e S_e^2(\kappa_1\epsilon_1 + \epsilon_2)$ and $n_e S_e^2(\kappa_1\epsilon_1 + \kappa_2\epsilon_2 + \epsilon_3)$. The upper bound of the significance level is then α.

Method (c)

Let $U_1, \ldots, U_{\kappa_1}; U_{\kappa_1+1}, \ldots, U_{\kappa_1+\kappa_2}$ be an orthonormal basis of the space formed by the vectors $Y_1, \ldots, Y_{\kappa_1}; Y_{\kappa_1+1}, \ldots, Y_{\kappa_1+\kappa_2}$, so that $U_1, \ldots, U_{\kappa_1}$ is a basis of the space of the vectors $Y_1, \ldots, Y_{\kappa_1}$. We shall show that when $\kappa_1 < \kappa_2$, it is possible to find orthogonal sets of κ_1 and $\kappa_2 - \kappa_1$ linear functions for the hypothesis H_1 and H_2, so that the inequality (5·3) could be used.

In (5·6), since the rank of β is at most κ_1, by a transformation of the vectors $\bar{Y}_1, \ldots, \bar{Y}_{\kappa_1}; \bar{Y}_{\kappa_1+1}, \ldots, \bar{Y}_{\kappa_1+\kappa_2}$, the matrix β can be reduced to the triangular form with $\kappa_2 - \kappa_1$ rows of zeros, i.e. the transformed vectors $Y'_1, \ldots, Y'_{\kappa_1}; Y'_{\kappa_1+1}, \ldots, Y'_{\kappa_1+\kappa_2}$ would be given by

$$\left.\begin{aligned}
Y'_{\kappa_1+1} &= \beta''_{1,1}U_1 + \beta''_{1,2}U_2 + \ldots + \beta''_{1,\kappa}U_{\kappa_1} + \gamma''_{1,1}U_{\kappa_1+1} + \ldots + \gamma''_{1,\kappa_2}U_{\kappa_1+\kappa_2}, \\
Y'_{\kappa_1+2} &= \beta''_{2,2}U_2 + \ldots + \beta''_{2,\kappa_1}U_{\kappa_1} + \gamma''_{2,1}U_{\kappa_1+1} + \ldots + \gamma''_{2,\kappa_2}U_{\kappa_1+\kappa_2}, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
Y'_{2\kappa_1+1} &= \gamma''_{\kappa_1+1,1}U_{\kappa_1+1} + \ldots + \gamma''_{\kappa_1+1,\kappa_2}U_{\kappa_1+\kappa_2}, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
Y'_{\kappa_1+\kappa_2} &= \gamma''_{\kappa_2,1}U_{\kappa_1+1} + \ldots + \gamma''_{\kappa_2,\kappa_2}U_{\kappa_1+\kappa_2}.
\end{aligned}\right\} \tag{5·20}$$

Let $W_1, \ldots, W_{\kappa_2-\kappa_1}$ form an orthonormal basis of the vector space of $Y'_{2\kappa_1+1}, \ldots, Y'_{\kappa_1+\kappa_2}$. We then have

$$E\{(W_i, y)\} = \sum_{j=i}^{\kappa_2} \xi_{1,j}\pi_{\kappa_1+j} = \Phi_{\kappa_1+i} \quad (i = 1, \ldots, \kappa_2 - \kappa_1). \tag{5·21}$$

The hypothesis H_2: $\pi_{\kappa_1+1} = 0, \ \ldots, \ \pi_{\kappa_1+\kappa_2} = 0,$

implies H_2': $\Phi_{\kappa_1+1} = 0, \ \ldots, \ \Phi_{\kappa_1+\kappa_2} = 0,$

since $\Phi_{\kappa_1+1}, \ldots, \Phi_{\kappa_1+\kappa_2}$ are linear functions of $\pi_{\kappa_1+1}, \ldots, \pi_{\kappa_1+\kappa_2}$ and have the rank $\kappa_2 - \kappa_1$. Now $\bar{H}_2'$ (i.e. rejection of H_2') implies $\bar{H}_2$ (i.e. rejection of H_2). We therefore test for the hypothesis H_2' and reject H_2 only when H_2' is rejected. A confidence region for the parameters $\pi_{\kappa_1+1}, \ldots, \pi_{\kappa_1+\kappa_2}$ is obtained from $\sum_{i=1}^{\kappa_2-\kappa_1} [(W_i, y) - \Phi_{\kappa_1+i}]^2$ which is distributed independently of

$$\sum_{i=1}^{\kappa_1} [(U_i, y) - \Phi_i]^2,$$

and these sums of squares may be used in equations (5·3) and (4·7).

6. Relative merits of methods (*a*), (*b*) and (*c*)

The method (*b*) obviously gives a much closer inequality than (*a*), although the latter has the additional advantage that it gives confidence intervals for all contrasts of the parameters in the hypotheses H_1 and H_2, in the set M_t. However, in most practical cases, the parameters entering into the hypotheses H_1 and H_2 relate to quite distinct characters, and thus there would not, usually, be any need to consider contrasts involving both sets of parameters.

In method (*c*) we use orthogonal sets of linear functions $Y_1, \ldots, Y_{\kappa_1}$; $W_1, \ldots, W_{\kappa_2-\kappa_1}$, and form χ^2's with κ_1 and $\kappa_2 - \kappa_1$ d.f. from these instead of using χ^2's with κ_1 and κ_2 d.f. from the linear functions $Y_1, \ldots, Y_{\kappa_1}$; $Y_{\kappa_1+1}, \ldots, Y_{\kappa_1+\kappa_2}$, which would have given quasi-independent tests for these hypotheses. The loss of degrees of freedom is serious when $\kappa_2 - \kappa_1$ is small, as the F-table shows. Thus the method (*c*) gives a good result only when $\kappa_2 - \kappa_1$ is not too small.

On the other hand, the method (*b*) gives a wide inequality unless κ_1 is small. It is not possible, however, to make a simple quantitative statement about the relative advantages of the methods (*b*) and (*c*) without detailed numerical calculations. One could, however, follow a tentative rule for the use of these methods:

κ_1	small $\kappa_2 - \kappa_1$ small	use (*b*),
κ_1	small $\kappa_2 - \kappa_1$ not small	use (*b*) or (*c*),
κ_1	not small $\kappa_2 - \kappa_1$ small	use (*b*) or (*c*),
κ_1	not small $\kappa_2 - \kappa_1$ not small	use (*c*).

The idea of this paper arose in the analysis of the data referred to in the text and in discussions with Dr Hughes Bryan and Dr B. G. Greenberg. The numerical details of the calculations will be published in a subsequent paper. The author wants to put on record his appreciation of helpful discussions with Dr Hughes Bryan and Dr B. G. Greenberg, of the University of North Carolina, and of help from the editorial board of *Biometrika* in improving the presentation of this paper.

REFERENCES

KIMBALL, A. W. (1951). On dependent tests of significance in the analysis of variance. *Ann. Math. Statist.* **22**, 600.

NANDI, H. K. (1951). On the analysis of variance test. *Bull. Calcutta Statist. Ass.* **3**, 103.

ROY, S. N. & BOSE, R. C. (1953). Simultaneous confidence interval estimation. *Ann. Math. Statist.* **24**, 513.

SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**, 87.

TUKEY, J. (1952). Allowances for various types of error rates. Unpublished invited address, Blacksburg meeting of the Institute of Mathematical Statistics, March 1952.

RANK ANALYSIS OF INCOMPLETE BLOCK DESIGNS

III. SOME LARGE-SAMPLE RESULTS ON ESTIMATION AND POWER FOR A METHOD OF PAIRED COMPARISONS*

By RALPH ALLAN BRADLEY

Virginia Agricultural Experiment Station of the Virginia Polytechnic Institute

I. Introduction

1·1. *Rank analysis of incomplete block designs*

A new method for paired comparisons was discussed in two recent papers (Bradley & Terry, 1952; Terry, Bradley & Davis, 1952). A mathematical model was postulated and tests of significance were developed. The procedures were considered as special cases of a rank analysis of incomplete block designs, and many of the concepts may be extended from the limited considerations of paired comparisons to ranking in incomplete block designs with two or more treatments in an incomplete block. The appropriateness of the model for paired comparisons has been discussed by Hopkins (1954) and also by Bradley (1954*a*). A large section of tables for a test of significance on the equality of treatment effects was included in the first reference cited and additional tables were prepared by the present author (1954*b*). For all of the tests considered in the above-referenced work, large-sample approximations to the sampling distributions of the statistics, under the conditions of the null hypotheses, have been given.

1·2. *Review of the method of paired comparisons*

It is necessary to summarize the mathematical model of the method of paired comparisons in order to outline the objectives of this paper.

We postulated true treatment ratings or parameters, $\pi_1, \ldots, \pi_t$, for t treatments in an experiment involving paired comparisons. It was assumed that every $\pi_i \geqslant 0$, and, for convenience, that† $\sum_i \pi_i = 1$. Further definition followed with the assumption that, when treatment i appears with treatment j in a block, the probability that treatment i obtains the higher rating (or a rank of 1) is $\pi_i/(\pi_i+\pi_j)$. Assuming independence in probability of treatment comparisons, we wrote the likelihood function in its general form as

$$L = \prod_i \pi_i^{a_i} \prod_{i<j} (\pi_i+\pi_j)^{-n}, \tag{1}$$

where

$$a_i = 2n(t-1) - \sum_j{}' \sum_{k=1}^{n} r_{ijk}. \tag{2}$$

r_{ijk} is the rank of treatment i in the comparison with treatment j in the kth of n repetitions of the paired comparisons design.

* This project was supported by funds from the Research and Marketing Act of 1946, under Contract with the Agricultural Research Service, United States Department of Agriculture.

† $\sum_i$ and $\prod_i$ will indicate respectively sums and products with $i = 1, \ldots, t$. $\sum_i{}'$ will mean that one value of i that appears in the argument of the summation is omitted. $\prod_{i<j}$ or $\prod_{i \neq j}$ represent products, $i = 1, \ldots, t, j = 1, \ldots, t, i<j$ or $i \neq j$ respectively. Departures from these conventions will be specified.

The method of maximum likelihood was used to obtain estimators, $p_1, \ldots, p_t$, of the parameters, $\pi_1, \ldots, \pi_t$, and likelihood ratio tests have been used throughout discussions based on the model set down. A general class of tests of the null hypothesis,

$$H_0: \quad \pi_i = 1/t \quad (i = 1, \ldots, t),$$

against the alternative hypothesis,

$$H_a: \quad \pi_i = \pi(h) \quad (h = 1, \ldots, m;\ i = s_{h-1}+1, \ldots, s_h),$$

where $s_0 = 0$, $s_m = t$ and $\sum_{h=1}^{m} (s_h - s_{h-1})\pi(h) = 1$, was formulated. These are, of course, tests of the indistinguishability of treatment effects on some attribute, perhaps colour or flavour, of the treatments. Two special cases were considered by Bradley & Terry (1952) and the specialization comes under H_a.

Case (i): The hypothesis H_a becomes

$$H_1: \quad \text{No } \pi_i \text{ is assumed equal to any } \pi_j \ (i \neq j);$$

that is, in H_a, $m = t$.

The normal equations resulting from the use of the method of maximum likelihood reduced to

$$\frac{a_i}{p_i} - n \sum_j{}' (p_i + p_j)^{-1} = 0 \quad (i = 1, \ldots, t) \tag{3}$$

and

$$\sum_i p_i = 1. \tag{4}$$

In this paper we shall be particularly interested in the statistic* $T = -2 \ln \lambda_1$, where λ_1 is the likelihood ratio for this special test. In this case

$$T = nt(t-1)\ln 2 - 2B_1 \ln 10, \tag{5}$$

with†

$$B_1 = n \sum_{i<j} \log (p_i + p_j) - \sum_i a_i \log p_i. \tag{6}$$

Case (ii): H_a becomes

$$H_2: \quad \pi_i = \pi \quad (i = 1, \ldots, s); \qquad \pi_i = (1 - s\pi)/(t - s) \quad (i = s+1, \ldots, t).$$

The general hypothesis H_a is thus restricted to the case in which $m = 2$. The estimator p of π was given as

$$p = \frac{ns(4t - s - 3) - 2 \sum_{i=1}^{s} \sum_j{}' \sum_{k=1}^{n} r_{ijk}}{ns(5st - 2t^2 - 6s + 3t) - 2(2s - t) \sum_{i=1}^{s} \sum_j{}' \sum_{k=1}^{n} r_{ijk}}. \tag{7}$$

A test statistic and an approximate test procedure were set forth in the reference. It will not be necessary to review these methods in view of the remarks that follow in § II.

Abelson & Bradley (1954) considered factorial arrangements of treatments imposed on the paired comparisons design. We shall not consider that situation in this paper.

1·3. *Objectives*

The objectives of this paper evolve from the need of considering the behaviour of the developed tests of significance and estimates of population parameters when the assumptions of the null hypotheses may not be true. We shall be interested in the power of the test

* log and ln indicate common and natural logarithms respectively.

† B_1 is the statistic tabulated for small samples. See Bradley & Terry (1952) and Bradley (1954*b*).

based on T in (5), in the reliability of the estimators p_i of π_i defined by (2) and (3), and in comparisons of the power of this test procedure with those of other possible procedures.

In the initial reference, it has been shown that T, in (5), under H_0, has the distribution of χ^2 with $(t-1)$ degrees of freedom for large samples. The difficulty to be expected in attempting an exact evaluation of the power of the test based on T for small samples, even for very restricted sets of alternative values of the parameters, was also discussed. Accordingly, we limit our objectives here to a large-sample evaluation and comparison of power functions and to the estimation of variances and covariances using large-sample results.

We shall show that Case (ii) yields the 'sign test', the properties of which are well known, and we may then limit our attention to Case (i).

II. Case (ii) and the sign test

2·1. *Case (ii)*

We refer to H_2: $\pi_i = \pi$ $(i = 1, \ldots, s)$; $\pi_i = (1-s\pi)/(t-s)$ $(i = s+1, \ldots, t)$ and the estimator p of π given in (7). It was not originally noted that the test procedure here reverts back to a binomial test.

Comparisons of treatments in the first group of s treatments yield contributions to the sums of ranks in (7) of 3. Then, if X is defined to be the number of times a treatment of the first group ranks above (obtains rank 1) a treatment of the second group,

$$\sum_{i=1}^{s} \sum_{j}{}' \sum_{k=1}^{n} r_{ijk} = \frac{3ns(s-1)}{2} + 2ns(t-s) - X. \tag{8}$$

Substitution of (8) in (7) yields

$$p = X/[(2s-t)X + ns(t-s)^2]. \tag{9}$$

Now, from the model for paired comparisons, the probability that a treatment i of the first group ranks above a treatment j of the second group in any of the n repetitions is

$$P(r_{ijk}=1) = \frac{\pi}{\pi + \dfrac{1-s\pi}{t-s}} = \frac{\pi(t-s)}{1+(t-2s)\pi}, \tag{10}$$

$$(i = 1, \ldots, s;\ j = s+1, \ldots, t;\ k = 1, \ldots, n).$$

When we substitute the estimator p of π given in (9) in the form (10), we obtain the estimated probability,

$$\text{Est}\, P(r_{ijk}=1) = X/[ns(t-s)]. \tag{11}$$

There are $ns(t-s)$ comparisons of treatments of the first group with treatments of the second group, and it is now apparent that this special case reduces to a consideration of $ns(t-s)$ binomial trials. The test procedures reduce to those of the binomial or sign test.

2·2. *The sign test*

The properties of the sign test have been thoroughly investigated. Hemelrijk (1952) notes that it is probably the oldest test in existence and refers to an application by Arbuthnot in 1710. The variance and approximate normality of X are discussed in elementary texts, and exact power function evaluations can be obtained for small samples from tables of the binomial distribution or of the incomplete beta function. The power of the sign test for large samples is based on the approximate normality of the statistic.

Dixon & Mood (1946) showed that the relative efficiency of the sign test in comparison with the t test under assumptions suitable for the valid application of the latter test is $2/\pi$. Later, Dixon (1953) prepared small-sample tables of power efficiencies that indicated that the sign test compares more favourably with the t-test than indicated by the asymptotic value, $2/\pi$.

In considering applications of Case (ii), it is preferable to go directly to the use of the sign test based on the $ns(t-s)$ comparisons of treatments in the first group with those of the second group. It will usually be sufficient to estimate $P(r_{ijk}=1)$ and not necessary to consider estimating π itself. If required, the variance of p, the estimator of π, may be obtained from the variance of X in an approximation through the use of usual formulae for the variance of a ratio.

We shall devote the remainder of this paper to a consideration of Case (i).

III. ESTIMATION

3·1. *Asymptotic distribution of* $(t-1)$ *maximum-likelihood estimators*

We shall require the large-sample distribution of the maximum-likelihood estimators, $p_1, \ldots, p_t$, of Case (i) and their asymptotic variances and covariances. These results will be of interest in themselves and useful in the development of subsequent sections of this paper.

Some new notation will assist in this discussion. Let

$$f(x,\pi) = \prod_i \pi_i^{x_i} \prod_{i<j} (\pi_i+\pi_j)^{-1}, \tag{12}$$

where we use x and π to represent vectors, $(x_1, \ldots, x_t)$ and $(\pi_1, \ldots, \pi_t)$. Now x_i is the number of times treatment i obtains a ranking of unity in a repetition of paired comparisons. If $x_{i(k)}$ is the observation on x_i in the kth of n repetitions, the association with a_i of (1) is

$$\sum_{k=1}^{n} x_{i(k)} = a_i \quad (i = 1, \ldots, t). \tag{13}$$

The likelihood function L in (1) is now $\prod_{k=1}^{n} f(x_{(k)}, \pi)$. It is also convenient to define

$$\left.\begin{aligned} \lambda_{ii} &= \frac{1}{\pi_i} \sum_h{}' \pi_h(\pi_i+\pi_h)^{-2} \quad (i = 1, \ldots, t), \\ \lambda_{ij} &= -(\pi_i+\pi_j)^{-2} \quad (i \neq j;\ i,j = 1, \ldots, t). \end{aligned}\right\} \tag{14}$$

We shall require the means, variances and covariances of $x_1, \ldots, x_t$. Let x_{ij} be an indicator variate with the value unity if treatment i ranks above treatment j and zero otherwise. Then $x_i = \sum_j{}' x_{ij}$. x_{ij} is a binomial variate with expectation $\pi_i(\pi_i+\pi_j)^{-1}$, and variance $\pi_i\pi_j(\pi_i+\pi_j)^{-2}$. The variates x_{ij} making up the sum x_i are independent in probability and it follows that

$$E(x_i) = \pi_i \sum_h{}' (\pi_i+\pi_h)^{-1} \quad (i = 1, \ldots, t) \tag{15}$$

$$V(x_i) = \pi_i \sum_h{}' \pi_h(\pi_i+\pi_h)^{-2} = \pi_i^2\lambda_{ii} \quad (i = 1, \ldots, t), \tag{16}$$

and

$$\operatorname{cov}(x_i, x_j) = -\pi_i\pi_j(\pi_i+\pi_j)^{-2} = \pi_i\pi_j\lambda_{ij} \quad (i \neq j;\ i,j = 1, \ldots, t). \tag{17}$$

The parameters $\pi_1, \ldots, \pi_t$ are not independent but subject to the restriction $\sum_i \pi_i = 1$. Accordingly we may regard $p_1, \ldots, p_{t-1}$ as maximum-likelihood estimators of the independent parameters, $\pi_1, \ldots, \pi_{t-1}$, taking $\pi_t = 1 - \sum_{i=1}^{t-1} \pi_i$. Then $\sqrt{n}(p_1 - \pi_1), \ldots, \sqrt{n}(p_{t-1} - \pi_{t-1})$ have a joint limiting normal distribution with zero means subject to the verification of certain regularity conditions. The required regularity conditions are quite well known and given, for example, by Cramér (1946, § 33·3, p. 500) and Chanda (1954), who fully states the conditions but for the continuous case. We have verified the conditions for the likelihood function L of (1) expressed in terms of $f(x, \pi)$ of (12), but these demonstrations are omitted for brevity.

The dispersion matrix of the joint limiting normal distribution of our $(t-1)$ estimators is the matrix $[\lambda'_{ij}]^{-1}$ with the definition of λ'_{ij} below. We need only note that

$$\frac{\partial \ln f}{\partial \pi_i} = \frac{x_i}{\pi_i} - \frac{x_t}{\pi_t} - \sum_h{}' (\pi_i + \pi_h)^{-1} + \sum_h{}' (\pi_t + \pi_h)^{-1} \quad (i = 1, \ldots, (t-1)), \tag{18}$$

and, from (15), that

$$E\left[\left(\frac{\partial \ln f}{\partial \pi_i}\right)\left(\frac{\partial \ln f}{\partial \pi_j}\right)\right] = \operatorname{cov}\left[\left(\frac{x_i}{\pi_i} - \frac{x_t}{\pi_t}\right)\left(\frac{x_j}{\pi_j} - \frac{x_t}{\pi_t}\right)\right] = \lambda'_{ij} \quad (i, j = 1, \ldots, (t-1)),$$

thereby defining λ'_{ij}. It follows from (16) and (17) that

$$\lambda'_{ij} = \lambda_{ij} - \lambda_{it} - \lambda_{jt} + \lambda_{tt}. \tag{19}$$

The matrix $[\lambda'_{ij}]$ is non-negative definite since it is a dispersion matrix and it is positive definite since $x_1, \ldots, x_{t-1}$, and hence $\frac{\partial \ln f}{\partial \pi_1}, \ldots, \frac{\partial \ln f}{\partial \pi_{t-1}}$, are free of linear restrictions.

The conclusion, with the established validity of the necessary conditions, is that

$\sqrt{n}(p_1 - \pi_1), \ldots, \sqrt{n}(p_{t-1} - \pi_{t-1})$ have for large samples the multivariate normal distribution with zero means and dispersion matrix $[\lambda'_{ij}]^{-1}$.

3·2. *Variances and covariances of estimators*

In the preceding section, the parameter π_t was considered to be a function of the remaining $(t-1)$ parameters. That process introduced asymmetry into the elements of the dispersion matrix, $[\lambda'_{ij}]^{-1}$. This lack of symmetry is essentially artificial and will now be removed.

The $t \times t$ matrix, $[\lambda_{ij}]$, is singular in view of the definitions (14) and the fact that

$$\pi_i \lambda_{it} + \sum_j{}' \pi_j \lambda_{ij} = 0. \tag{20}$$

Then, if the elements of the last row and then of the last column of the matrix $[\lambda_{ij}]$ are subtracted from corresponding elements of the remaining rows and columns respectively,

$$|\lambda_{ij}| = \begin{vmatrix} [\lambda'_{ij}] & [\lambda_{it} - \lambda_{tt}]' \\ [\lambda_{jt} - \lambda_{tt}] & \lambda_{tt} \end{vmatrix} = 0,$$

where $[\lambda_{it} - \lambda_{tt}]'$ and $[\lambda_{jt} - \lambda_{tt}]$ are respectively column and row vectors of $(t-1)$ elements. It follows that

$$|\lambda'_{ij}| = \begin{vmatrix} [\lambda'_{ij}] & [\lambda_{it} - \lambda_{tt}]' \\ [\lambda_{jt} - \lambda_{tt}] & 1 + \lambda_{tt} \end{vmatrix}. \tag{21}$$

Now, by reversing the process and adding the elements of the last row and then the last column to corresponding elements of the remaining rows and columns in (21), we have

$$|\lambda'_{ij}| = |\lambda_{ij}+1| = \begin{vmatrix} [\lambda_{ij}+1] & [1]' \\ [0] & 1 \end{vmatrix} = -\begin{vmatrix} [\lambda_{ij}] & [1]' \\ [1] & 0 \end{vmatrix}, \tag{22}$$

where [1] and [1]′ are respectively row and column vectors of t unit elements. The last step again depends on the result that $|\lambda_{ij}| = 0$. In the same way it can be shown that the cofactor of λ_{ij} in the extreme right-hand side of (22) is equal to the cofactor of λ'_{ij} in $|\lambda'_{ij}|$. Then, if σ_{ij} is the covariance of $\sqrt{n}(p_i-\pi_i)$ and $\sqrt{n}(p_j-\pi_j)$ $(i,j = 1,\ldots,(t-1))$,

$$\sigma_{ij} = \frac{\text{cofactor of } \lambda_{ij} \text{ in } \begin{vmatrix} [\lambda_{ij}] & [1]' \\ [1] & 0 \end{vmatrix}}{\begin{vmatrix} [\lambda_{ij}] & [1]' \\ [1] & 0 \end{vmatrix}}. \tag{23}$$

But the formulation of the model for paired comparisons is symmetric in the parameters $\pi_1,\ldots,\pi_t$ and the estimators $p_1,\ldots,p_t$, and (23) applies for all variances and covariances with $i,j = 1,\ldots,t$ on the basis of its symmetry.

The variances and covariances obtainable from (23) are simply the elements of the t-square principal minor of the inverse of the matrix of the determinant in the denominator of (23). For small values of t, this inverse matrix may be evaluated by elementary methods when $\pi_1,\ldots,\pi_t$ are specified or replaced by estimates. However, since $|\lambda_{ij}| = 0$, the usual Doolittle methods of matrix inversion break down. Then for larger values of t, it appears desirable to invert $[\lambda'_{ij}]$ specified through (19) using a Doolittle method, thus obtaining the required variances and covariances with $i,j = 1,\ldots,(t-1)$. The remaining variance and covariances associated with $\sqrt{n}(p_t-\pi_t)$ are determined through the relationship

$$\sqrt{n}(p_t-\pi_t) = -\sum_{i=1}^{t-1}\sqrt{n}(p_i-\pi_i),$$

so that $$\sigma_{tt} = \sum_{i=1}^{t-1}\sum_{j=1}^{t-1}\sigma_{ij} \quad \text{and} \quad \sigma_{it} = -\sum_{j=1}^{t-1}\sigma_{ij} \quad (i = 1,\ldots,(t-1)).$$

Since it has been established in § 3·1 that $\sqrt{n}(p_1-\pi_1),\ldots,\sqrt{n}(p_{t-1}-\pi_{t-1})$ have asymptotically a multivariate normal distribution, and since $\sqrt{n}(p_t-\pi_t)$ is a linear function of those variates, we may now state that

$\sqrt{n}(p_1-\pi_1),\ldots,\sqrt{n}(p_t-\pi_t)$ have, for large values of n, the singular multivariate normal distribution of $(t-1)$ dimensions in a space of t dimensions with zero means and dispersion matrix, $[\sigma_{ij}]$, defined through (23).

In general, for large samples, we may take $p_1,\ldots,p_t$ to be jointly normally distributed with means $\pi_1,\ldots,\pi_t$, and dispersion matrix $[\sigma_{ij}]/n$. Then any linear function $\sum_i b_i p_i$ may be taken to be normal for large samples with mean $\sum_i b_i\pi_i$ and variance $\sum_{i<j} b_i b_j \sigma_{ij}/n$. In particular, we may be interested in orthogonal linear comparisons of the sort often considered in the analysis of variance.

Estimated variances and covariances will usually be required. Through the consistency of maximum-likelihood estimators, we can define $\hat{\lambda}_{ij}$ $(i,j = 1,\ldots,t)$ to be the same functions of $p_1,\ldots,p_t$ as λ_{ij} are of $\pi_1,\ldots,\pi_t$ as defined in (14). Then $[\hat{\sigma}_{ij}]$, the sample dispersion matrix, is the same function of $\hat{\lambda}_{ij}$ as $[\sigma_{ij}]$ is of λ_{ij} in definition (23).

It has been shown (Bradley, 1953) that the functions of the parameters, $\ln \pi_1, \ldots, \ln \pi_t$, determine in a sense location points for the t treatments on an arithmetic scale. Consequently, there may be some interest in obtaining

$$\sigma_{ij}(\ln p_i, \ln p_j) \approx \frac{\sigma_{ij}}{n\pi_i\pi_j} \quad (i,j = 1, \ldots, t), \tag{24}$$

as given, for example, by Hald (1952). If common logarithms are used, a factor, $(0{\cdot}4343)^2$, is required in the right-hand member of (24).

IV. Large-sample distributions of T

4·1. *T expanded in a series*

It has been shown (Bradley & Terry, 1952) that, under the conditions of the null hypothesis of Case (i) of § 1·2, T as defined in (5) has, asymptotically with n, the χ^2-distribution with $(t-1)$ degrees of freedom. In order to investigate the power of the test procedure based on T and to consider the efficiency of the method in comparison with other test procedures, we require the distribution of T under the alternative hypothesis H_a of Case (i). In this section we shall show that T has the same limiting distribution as a certain sum of squares.

Let
$$y_i = t\left(p_i - \frac{1}{t}\right) \quad (i = 1, \ldots, t), \tag{25}$$

and note that from (2)
$$a_i = \tfrac{1}{2}n(1+y_i)\sum_j{}' [1 + \tfrac{1}{2}(y_i+y_j)]^{-1} \tag{26}$$

for all i, and
$$\sum_i a_i = \tfrac{1}{2}nt(t-1). \tag{27}$$

Substitution in (5) through the use of (6) and upon reduction based on (26) and (27) yields

$$T = n\sum_i (1+y_i)[\ln(1+y_i)]\sum_j{}' [1+\tfrac{1}{2}(y_i+y_j)]^{-1} - 2n\sum_{i<j}\ln[1+\tfrac{1}{2}(y_i+y_j)]^{-1}. \tag{28}$$

It is next possible, by expanding the right-hand member of (28) in a power series in the variates, to show that
$$T = \tfrac{1}{4}nt\sum_i y_i^2 + R(y_i), \tag{29}$$

where $R(y_i)$ depends on higher powers of the variates than the second. In turn, it can be shown that

$$|R(y_i)| \leqslant \tfrac{1}{12}n(5t-8)\sum_i |y_i|^3 + \tfrac{1}{12}n(9t-14)\sum_i y_i^4 + \tfrac{1}{8}n(t-2)\sum_i |y_i|^5 + \tfrac{1}{12}n\sum_{i<j}|y_i+y_j|^3$$
$$+ n\sum_i [|y_i| + \tfrac{1}{2}y_i^2 + \tfrac{5}{6}|y_i|^3 + \tfrac{1}{3}y_i^4]\sum_j{}' \frac{(y_i+y_j)^2}{4[1+\tfrac{1}{2}(y_i+y_j)]}. \tag{30}$$

This result, (30), is obtained through algebraic manipulation involving the use of the general relations
$$|\ln(1+x) - x + \tfrac{1}{2}x^2| \leqslant |\tfrac{1}{3}x|^3, \quad (1+x)^{-1} = 1 - x + x^2/(1+x),$$

and the particular relations

$$\sum_j{}' y_j = -y_i \quad \text{and} \quad 2\sum_{i<j} y_i y_j = -\sum_i y_i^2.$$

Let us suppose that $\sqrt{n}\,y_i$ $(i = 1, \ldots, t)$ have limiting distributions and this will be demonstrated under desired conditions. Assuming this result, we may then state that $n^{\frac{1}{2}-\epsilon}y_i$, for any $\epsilon > 0$, converges in probability to zero, for it is an easily proved theorem in probability that, if (X_N) represents a sequence of random variables with a limiting distribution function and if (α_N) is a sequence of constants approaching zero as $N \to \infty$, then $(\alpha_N X_N)$ is a sequence of random variables converging in probability to zero as $N \to \infty$. It follows from Slutsky's theorem (Cramér, 1946, § 20·6, p. 255) that, if $\sqrt{n}\,y_i$ have limiting distributions, $R(y_i)$ converges stochastically to zero. *This is sufficient* (Cramér, 1946, § 20·6, p. 254) *to state that* T *has the same limiting distribution as* $\frac{1}{4}nt\sum_i y_i^2$.

4·2. *Limiting distribution of* y_i *under* H_a

The normal equations for the maximum-likelihood estimators were given in (3) and (4), and these relations are useful in obtaining the limiting distribution of $\sqrt{n}\,y_i$ defined in (25) under modified specifications of $\pi_1, \ldots, \pi_t$. From (3) or (26), we have

$$a_i/n = \tfrac{1}{2}(1+y_i)\sum_j{}' [1+\tfrac{1}{2}(y_i+y_j)]^{-1} \quad (i = 1, \ldots, t). \tag{31}$$

Expansion in (31), using the negative binomial and the result that $\sum_j{}' y_j = -y_i$, and multiplication by $\sqrt{n}$ lets us write

$$\sqrt{n}\left[\frac{a_i}{n} - \tfrac{1}{2}(t-1)\right] = \tfrac{1}{4}t\sqrt{n}\,y_i - \frac{(t-2)}{4\sqrt{n}}(\sqrt{n}\,y_i)^2 + \frac{1}{8\sqrt{n}}\sum_j{}' \frac{(1+y_i)(\sqrt{n}\,y_i+\sqrt{n}\,y_j)^2}{[1+\tfrac{1}{2}(y_i+y_j)]}. \tag{32}$$

Now, from an argument similar to that used in § 4·1, if $\sqrt{n}\,y_i$ has a limiting distribution, it has the same limiting distribution as

$$\frac{4\sqrt{n}}{t}\left[\frac{a_i}{n} - \tfrac{1}{2}(t-1)\right]. \tag{33}$$

Let us now redefine the parameters so that

$$\pi_i = \frac{1}{t} + \frac{\delta_{in}}{\sqrt{n}} \quad (i = 1, \ldots, t), \tag{34}$$

where δ_{in} represents a sequence of constants converging to δ_i as $n \to \infty$. This redefinition of the parameters is an artifice that permits us to find the limiting distribution of T under H_a and thence the power of the test procedure. This essentially means that we are investigating the asymptotic power for alternatives in the locality of the parameter point determined by the null hypothesis. This is necessary in order that we do not merely show that the test procedure is consistent. Under these conditions we require the limiting distribution of the variate defined in (33).

Let $\bar{x}_{ij}$ be the average over the n repetitions of the binomial variate, x_{ij}, defined in § 3·1. Then, through the definitions of x_i and $x_{i(k)}$ and (13), we have

$$a_i = n\sum_j{}' \bar{x}_{ij}, \tag{35}$$

and the variates, $\bar{x}_{i1}, \ldots, \bar{x}_{it}$, omitting $\bar{x}_{ii}$, are independent.

Since $E(x_{ij}) = \pi_i(\pi_i+\pi_j)^{-1}$, we may define $\phi_n(\tau)$ to be the characteristic function of $2\sqrt{n}\,(\bar{x}_{ij} - \frac{1}{2})$, and, following the method of Cramér (1946) in § 16·4 of his book, we have

$$\phi_n(\tau) = \left[\frac{\pi_j \exp\left\{-\sqrt{-1}\,\frac{\tau}{\sqrt{n}}\right\}}{(\pi_i+\pi_j)} + \frac{\pi_i \exp\left\{\sqrt{-1}\,\frac{\tau}{\sqrt{n}}\right\}}{(\pi_i+\pi_j)}\right]^n. \tag{36}$$

Expansion of the exponentials in (36) yields

$$\phi_n(\tau) = \left[1 + \left(\frac{\pi_i - \pi_j}{\pi_i + \pi_j}\right)\sqrt{-1}\,\frac{\tau}{\sqrt{n}} - \frac{\tau^2}{2n} + \left(\frac{\pi_j\gamma^3 - \pi_i\theta^3}{\pi_i + \pi_j}\right)\frac{\tau^3}{6n^{\frac{3}{2}}}\right]^n, \tag{37}$$

where γ and θ are real or complex quantities less than unity in modulus.

We now substitute from (34) for the parameters in (37) and write

$$\phi_n(\tau) = \left[1 + \sqrt{-1}\,\frac{t\tau}{2n}(\delta_i - \delta_j) + \frac{H}{n} - \frac{\tau^2}{2n}(1+W)\right]^n,$$

where $$H = \sqrt{-1}\,\frac{t\tau}{2}(\delta_{in} - \delta_i - \delta_{jn} + \delta_j) - \sqrt{-1}\,\frac{t^2\tau}{4\sqrt{n}}(\delta_{in}^2 - \delta_{jn}^2)\left[1 + \frac{t}{2\sqrt{n}}(\delta_{in} + \delta_{jn})\right]^{-1},$$

and $$W = \frac{\tau}{3\sqrt{n}}\left[\frac{1}{t}(t^3 - \gamma^3) + \frac{\delta_{in}\theta^3}{\sqrt{n}} - \frac{\delta_{jn}\gamma^3}{\sqrt{n}}\right]\left[\frac{2}{t} + \frac{\delta_{in} + \delta_{jn}}{\sqrt{n}}\right]^{-1}.$$

From the convergence of δ_{in} to δ_i and the forms of H and W, for any ϵ, $\eta > 0$, there exists $n_0(\epsilon, \eta)$ such that, when $n > n_0$, $|H| < \epsilon$ and $|W| < \eta$ for fixed τ. It follows that

$$\exp[-\tfrac{1}{2}\tau^2(1+\eta) + \tfrac{1}{2}\sqrt{-1}t\tau(\delta_i - \delta_j) - \epsilon] \leqslant \lim_{n\to\infty}\phi_n(\tau) \leqslant \exp[-\tfrac{1}{2}\tau^2(1-\eta) + \tfrac{1}{2}\sqrt{-1}t\tau(\delta_i - \delta_j) + \epsilon],$$

and hence $$\lim_{n\to\infty}\phi_n(\tau) = \exp[-\tfrac{1}{2}\tau^2 + \tfrac{1}{2}\sqrt{-1}t\tau(\delta_i - \delta_j)]. \tag{38}$$

This result (38) is sufficient (see, for example, Cramér, 1946, § 10·4) to state that $2\sqrt{n}(x_{ij} - \frac{1}{2})$ has a normal limiting distribution with mean $\frac{1}{2}t(\delta_i - \delta_j)$ and unit variance.

We may obtain the variances and covariances of $4\sqrt{n}[a_i/n - \frac{1}{2}(t-1)]/t$ through use of the relation (35). It is easy to show that

$$\frac{4\sqrt{n}}{t}\left[\frac{a_i}{n} - \tfrac{1}{2}(t-1)\right] = \frac{2}{t}\sum_j{}'[2\sqrt{n}(\bar{x}_{ij} - \tfrac{1}{2})], \tag{39}$$

and we note also that by definition $$\bar{x}_{ji} = 1 - \bar{x}_{ij}. \tag{40}$$

The mean of the left-hand member of (39) in the limit is now seen to be $\sum_j{}'(\delta_i - \delta_j) = t\delta_i$, since $\sum_j{}'\delta_j = -\delta_i$. Through the variance of $2\sqrt{n}(\bar{x}_{ij} - \frac{1}{2})$ and the independence of the variates in the sum in (39), the limiting variance of the left-hand member of (39) is $4(t-1)/t^2$. $\bar{x}_{ij}$ and $\bar{x}_{gh}$ are independent unless $i = g$ and $j = h$ or $i = h$ and $j = g$. Then in view of (40), the covariance of $4\sqrt{n}[a_i/n - \frac{1}{2}(t-1)]/t$, and $4\sqrt{n}[a_j/n - \frac{1}{2}(t-1)]/t$ $(i \neq j)$ is $-4/t^2$. *It follows that* $4\sqrt{n}[a_1/n - \frac{1}{2}(t-1)]/t$, ..., $4\sqrt{n}[a_t - \frac{1}{2}(t-1)]/t$, *and consequently* $\sqrt{n}\,y_1$, ..., $\sqrt{n}\,y_t$, *have a joint limiting normal distribution with means* $t\delta_1$, ..., $t\delta_t$, *and equal variances and covariances* $4(t-1)/t^2$ *and* $-4/t^2$, *respectively under the conditions* (34) *on the parameters as* $n \to \infty$. Since $\sum_i y_i = 0$, the limiting distribution is necessarily singular.

It is to be noted that this same limiting distribution can apparently be obtained by using the definition (34) for the parameters and relying on the use of the joint limiting distribution of $\sqrt{n}(p_1 - \pi_1)$, ..., $\sqrt{n}(p_t - \pi_t)$ developed in § 3·2 above. If π_i is replaced by $t^{-1} + n^{-\frac{1}{2}}\delta_{in}$ in the definitions of λ_{ij} in (14), and if λ_{ij} is then replaced by $\lim_{n\to\infty}\lambda_{ij}$ in (23), the correct variances and covariances would be obtained for $\sqrt{n}\,y_1$, ..., $\sqrt{n}\,y_t$. That this procedure is valid has been proved by Wald (1943) but we have obtained the results directly.

4·3. *Limiting distribution of T under H_a: $\pi_i = t^{-1} + n^{-\frac{1}{2}}\delta_{in}$.*

We define
$$u_i = \tfrac{1}{2}\sqrt{t}\sqrt{n}\,y_i, \tag{41}$$
and, from § 4·1, T has the same limiting distribution as $\sum_i u_i^2$. $u_1, \ldots, u_t$ have a joint limiting distribution that is multivariate normal with means, $\frac{1}{2}t^{\frac{1}{2}}\delta_1, \ldots, \frac{1}{2}t^{\frac{1}{2}}\delta_t$, variances $(t-1)/t$ and covariances $-1/t$. Further, $\sum_i u_i = 0$.

The Helmert transformation is used to transform to new variables, $z_1, \ldots, z_t$, with $z_t = 0$, and hence T has the same limiting distribution as $\sum_{i=1}^{t-1} z_i^2$. It is easily verified that $z_1, \ldots, z_{t-1}$ are independent, have unit variances, and have limiting normal distributions under H_a with means, say, $\theta_1, \ldots, \theta_{t-1}$, where $\sum_{i=1}^{t-1}\theta_i^2 = \frac{1}{4}t^3\sum_i \delta_i^2$. This is sufficient to permit the final conclusion that $\sum_{i=1}^{t-1} z_i^2$ *and T have a limiting distribution under H_a: $\pi_i = t^{-1} + n^{-\frac{1}{2}}\delta_{in}$ as $n \to \infty$, that is, the non-central χ^2-distribution with $(t-1)$ degrees of freedom and parameter of non-centrality,* $\lambda = \frac{1}{4}t^3\sum_i \delta_i^2$. λ is defined if we reiterate that the limiting distribution function of T under H_a is obtained by integrating the non-central χ^2-density,
$$f(T) = \frac{e^{-\frac{1}{2}T}\, e^{-\frac{1}{2}\lambda}}{2^{\frac{1}{2}(t-1)}} \sum_{h=0}^{\infty} \frac{T^{\frac{1}{2}(t-1)+h-1}\lambda^h}{\Gamma[\frac{1}{2}(t-1)+h]\, 2^{2h} h!}. \tag{42}$$

The power of the paired comparisons test based on T is asymptotically given by
$$\beta(\lambda \mid \alpha, t-1, \infty) = \int_{\chi^2_{\alpha,t-1}}^{\infty} f(T)\, dT, \tag{43}$$
where $\chi^2_{\alpha,t-1}$ is the α-level significance value of a central χ^2-distribution with $(t-1)$ degrees of freedom. Fix (1949) provided tables for $\alpha = 0{\cdot}01$ and $\alpha = 0{\cdot}05$ with λ tabulated for given values of β and for given degrees of freedom. Pearson & Hartley (1951) provided charts for given degrees of freedom and $\alpha = 0{\cdot}01$ and $0{\cdot}05$, whereon β is plotted against ϕ. These charts are for the non-central F-distribution but may be used for the non-central χ^2-distribution if the denominator degrees of freedom of F are taken to be infinite. The notation $\beta(\lambda \mid \alpha, \nu_1, \nu_2)$ was used by Pearson & Hartley and has been modified for our special case in (43).
$$\phi = \sqrt{(\lambda/t)} = \tfrac{1}{2}t\sqrt{\textstyle\sum_i \delta_i^2}. \tag{44}$$

Tang (1938) presented tables based on ϕ that may also be used to evaluate the integral (43).

For large n, an adequate approximation to the power of the paired comparisons test may be obtained by assuming the non-central χ^2-distribution and taking the parameter λ to be
$$\lambda = \tfrac{1}{4}nt^3\sum_i (\pi_i - 1/t)^2. \tag{45}$$

When H_0 is true, $\pi_i = 1/t$ and $\delta_{in} = \delta_i = 0$ $(i = 1, \ldots, t)$. Accordingly, $\lambda = 0$ and T has then the central χ^2-distribution with $(t-1)$ degrees of freedom for large n. This simply re-establishes the approximate large-sample distribution for the test of H_0 given by Bradley & Terry (1952).

V. Comparisons of powers of tests

5·1. *Comparison of the test for paired comparisons with a multi-binomial test*

One of the chief uses of an investigation of the power of a statistical test is an assessment of the merits of the test in contrast with alternative procedures. In this and the following section we shall compare the test based on T with two other possible methods.

The model formulated for the method of paired comparisons is not the most general one of its form possible. We could postulate parameters π_{ij}, the probability of treatment i being rated above treatment j in the comparison of these two treatments. Then the n repetitions of the paired comparisons design could be regarded as $\frac{1}{2}t(t-1)$ unrelated binomial tests, each test depending on n trials.

Let us consider the comparison of treatment i with treatment j. The usual test of the null hypothesis, $\pi_{ij} = \frac{1}{2}$, is based on the statistic,

$$S_{ij} = 4n(p_{ij} - \tfrac{1}{2})^2, \tag{46}$$

where p_{ij} is the estimator of π_{ij}. Further,

$$S_{ij} = 2n(p_{ij} - \tfrac{1}{2})^2 + 2n(p_{ji} - \tfrac{1}{2})^2 = u_{ij}^2 + u_{ji}^2$$

and u_{ij} and u_{ji} correspond to variates u_i defined in (41). This binomial test is only a special case, $t = 2$, of the paired comparisons method being discussed in this paper. Then, under the null hypothesis, S_{ij} has a limiting χ^2-distribution with 1 degree of freedom. (This result is of course well known.) Further, if the alternative hypothesis is expressed as

$$\pi_{ij} = \tfrac{1}{2} + \mu_{ijn}/\sqrt{n}$$

with μ_{ijn} converging to μ_{ij} as $n \to \infty$, S_{ij} has a limiting non-central χ^2-distribution with 1 degree of freedom and parameter of non-centrality, $\lambda_{ij} = 2(\mu_{ij}^2 + \mu_{ji}^2)$, $\pi_{ij} = 1 - \pi_{ji}$ and $\mu_{ij} = -\mu_{ji}$. The distribution of S_{ij} under the alternative hypothesis follows as a special case of the theory developed for paired comparisons.

Suppose that the model for paired comparisons is appropriate. Then, $\pi_{ij} = \pi_i/(\pi_i + \pi_j)$ and, if $\pi_i = t^{-1} + n^{-\frac{1}{2}}\delta_{in}$ as before,

$$\mu_{ijn} = \tfrac{1}{4}t(\delta_{in} - \delta_{jn}) - \frac{t^2}{4\sqrt{n}}\delta_{in}(\delta_{in} + \delta_{jn}) + \frac{t^3}{8\sqrt{n}}\left(\frac{1}{t} + \frac{\delta_{in}}{\sqrt{n}}\right)(\delta_{in} + \delta_{jn})^2\left[1 + \frac{t}{2\sqrt{n}}(\delta_{in} + \delta_{jn})\right]^{-1}.$$

It follows at once that

$$\mu_{ij} = \tfrac{1}{4}t(\delta_i - \delta_j), \tag{47}$$

and

$$\lambda_{ij} = \tfrac{1}{4}t^2(\delta_i - \delta_j)^2. \tag{48}$$

If an overall test of H_0, the equality of all treatment parameters, here called the multi-binomial test, were made on the basis of the $\frac{1}{2}t(t-1)$ independent sets of binomial trials, the appropriate statistic would be

$$S = \sum_{i<j} S_{ij}. \tag{49}$$

From the foregoing argument and the additive property of the χ^2-distribution, under H_0, S has the central χ^2-distribution with $\frac{1}{2}t(t-1)$ degrees of freedom and, under H_a, S has the non-central χ^2-distribution with the same degrees of freedom and parameter of non-centrality,

$$\lambda' = \tfrac{1}{4}t^2\sum_{i<j}(\delta_i - \delta_j)^2 = \tfrac{1}{4}t^2\sum_i \delta_i^2 = \lambda, \tag{50}$$

in the limit with n. The simplification in (50) is possible since $\sum_i \delta_i = 0$.

The paired comparisons test and the multi-binomial test under the conditions appropriate to the former have asymptotic power functions that differ only in the numbers of degrees of freedom that are available. The asymptotic powers of the two tests are clearly different. In Figs. 1 and 2 we have plotted the power β against λ for $t = 3$ and $t = 4$ for both tests based on both 0·05 and 0·01 levels of significance. The β-scale is logarithmic and the power curves were obtained from similar charts given by Pearson & Hartley (1951). The horizontal scales on our charts are for λ, the element common to the parameters of the test considered. To use the Pearson & Hartley charts we computed $\phi = \sqrt{(\lambda/t)}$ for the paired comparisons test as given in (44) and we required

$$\phi' = \sqrt{\frac{\lambda}{\frac{1}{2}t(t-1)+1}} \tag{51}$$

for the multi-binomial test. It is to be noted that, when $\lambda = 0$, $\beta = 0{\cdot}05$ or $0{\cdot}01$ for both tests. Further, as one would expect, $\beta \to 1$ as $\lambda \to \infty$ for both tests. In the intervening λ-region, the T-test is more powerful than the multi-binomial test, and there is an indication that the advantage of the paired comparisons test increases sharply with t, the number of treatments.

Power curves for an analysis-of-variance test are also given in Figs. 1 and 2. The comparison of paired comparisons with analysis of variance is discussed in the next section.

5·2. *Comparison of the test for paired comparisons with the analysis of variance*

A correspondence between the parameters of the method of paired comparisons and those of analysis of variance may be established by referring to the discussion of Thurstone's model for paired comparisons given by Mosteller (1951). Mosteller summarizes the Thurstone model by writing*

$$P(X_i > X_j) = \frac{1}{\sqrt{(2\pi)}} \int_{-(S_i - S_j)/\sigma_d}^{\infty} e^{-\frac{1}{2}v^2}\, dy, \tag{52}$$

where S_i is the location point for the variate X_i on a subjective continuum and σ_d^2 is the variance of the difference $(X_i - X_j)$. It was assumed that the variance of X_i is σ^2 and the covariance of X_i and X_j is $\rho\sigma^2$. Then $\sigma_d^2 = 2\sigma^2(1-\rho)$. If X_i and X_j can be observed in paired comparisons, the usual additive model of the analysis of variance would require that $\rho = 0$, S_i be the effect of treatment i, and σ^2 be the common variance of the random elements of that additive model. For our purposes, the Thurstone model may more generally be associated with any analysis of variance with the same variance σ^2 for the random element of its additive model and with S_i as the effect of treatment i.

The present author (1953) has shown that, for the method of paired comparisons here discussed, one can write

$$P(X_i > X_j) = \frac{\pi_i}{\pi_i + \pi_j} = \frac{1}{4}\int_{-(\ln \pi_i - \ln \pi_j)}^{\infty} \operatorname{sech}^2 \tfrac{1}{2}y\, dy = \frac{1}{4a}\int_{-a(\ln \pi_i - \ln \pi_j)}^{\infty} \operatorname{sech}^2 \frac{y}{2a}\, dy. \tag{53}$$

A correspondence between S_i and π_i is required to compare the method of paired comparisons with the analysis of variance. That the forms of the integrand in (52) and (53) do not matter for large-sample sizes can be demonstrated.

* Mosteller indicated in his discussion that no generality was lost by taking $\sigma_d = 1$ and he wrote (52) with that restriction.

Consider

$$P(X_i > X_j) = \int_{-c}^{\infty} f(y)\,dy.$$

The variance of an estimate c_e of c, obtained from an estimate p of $P(X_i > X_j)$, can for large samples be written*

$$V(c_e) = \frac{1}{n} P(X_i > X_j)\{1 - P(X_i > X_j)\}\{f(-c)\}^{-2}.$$

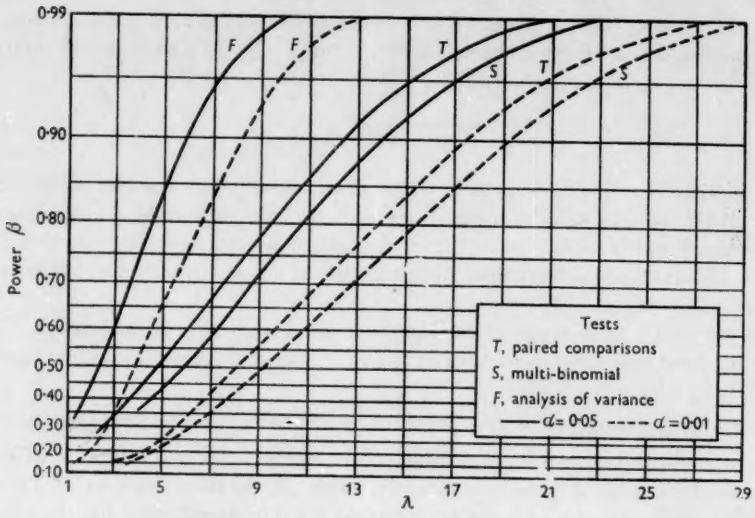


Fig. 1. Asymptotic power functions: $t = 3$.

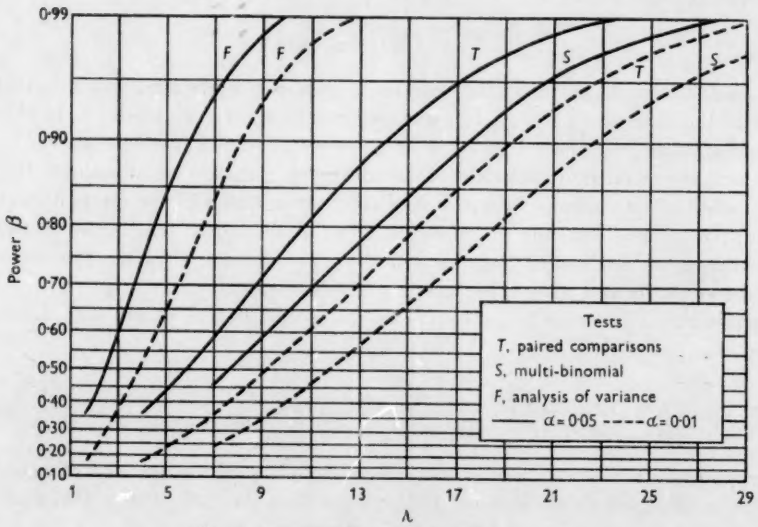


Fig. 2. Asymptotic power functions: $t = 4$.

* For a derivation when $f(y)$ is normal, see Cochran (1937). The result is used by Dixon & Mood (1946) in discussing the statistical sign test.

$V(c_e)$ does not depend on the form of $f(y)$ but only on the ordinate $f(-c)$. Accordingly, we establish a correspondence between S_i and π_i when we adjust the scale factor, a, in (53) so that the ordinate is equal to $1/\sqrt{(2\pi)}$. The ordinate in both (52) and (53) is evaluated at $y = 0$ since $\ln \pi_i - \ln \pi_j = O(n^{-\frac{1}{2}})$ because of the definition, $\pi_i = t^{-1} + n^{-\frac{1}{2}}\delta_{in}$. It follows that

$$a = \sqrt{(\tfrac{1}{8}\pi)} \tag{54}$$

and $\sqrt{(\frac{1}{8}\pi)} \ln \pi_i$ corresponds to S_i/σ_d.

If in the Thurstone model we assume that $\rho = 0$ as required for the association with analysis of variance, $\sqrt{2}\, S_i/\sigma_d$ represents the location point (or treatment effect) for the ith treatment on a subjective continuum (or additive scale) whereon the variance of an observation is unity. In terms of the parameters of the method of paired comparisons, the location point is $\sqrt{(\frac{1}{4}\pi)} \ln \pi_i$. If $n(t-1)$ observations are made on each variate X_i in the analysis of variance (this corresponds to the number of times a treatment appears in the method of paired comparisons), the parameter of non-centrality for the F-test would be approximately

$$\lambda''_n = \frac{\pi n(t-1)}{4} \sum_i \left(\ln \pi_i - \frac{1}{t} \sum_j \ln \pi_j \right)^2. \tag{55}$$

For large samples the non-central F-distribution approaches the distribution of the non-central χ^2 with $(t-1)$ degrees of freedom and $\lambda''_n \to \lambda''$, where

$$\lambda'' = \tfrac{1}{4}\pi t^2 (t-1) \sum_i \delta_i^2 = \pi(t-1)\lambda/t, \tag{56}$$

in view of (34) and (50).

The power functions for analysis of variance are plotted in Figs. 1 and 2. To plot these curves, we computed $\phi'' = \sqrt{[\pi(t-1)\lambda/t^2]}$ in order to use the Pearson & Hartley charts. It is clear that analysis of variance is superior to paired comparisons and the multi-binomial procedure for $t = 3$ and 4 and the advantage increases with t. On the other hand, for comparative purpose, we have assumed that conditions for the valid application of analysis of variance could be attained while the method of paired comparisons was devised for situations where the analysis of variance may not be used.

5·3. *Relative efficiencies*

The relative efficiency of one test to another may be obtained from the limiting ratio of sample sizes required for equal powers. The concept depends on local properties of the power functions of the two tests being compared. Relative efficiency is the limit of the inverse ratio of sample sizes required in two tests in order that the tests have equal powers for alternatives approaching, with increasing sample sizes, parameter values specified by the null hypothesis. The methods employed in this section parallel those developed by Pitman (1948). We shall first obtain the asymptotic relative efficiency of paired comparisons to analysis of variance by elementary means. Then, to complete the inter-comparisons, we use a more general theorem of Pitman, as presented and extended by Noether (1955), to obtain the relative efficiencies of the multi-binomial procedure to paired comparisons and to the analysis of variance.

Let n'' be the same size for the analysis of variance and n the sample size for paired comparisons. For comparable treatment differences with the ratio n''/n fixed, we need $\delta_{in''}$ corresponding to δ_{in} of (34) such that

$$\lim_{n,\, n'' \to \infty} \left\{ \sum_i \frac{\delta''^2_{in''}}{n''} \Big/ \sum_i \frac{\delta_{in}^2}{n} \right\} = 1. \tag{57}$$

Now we have already seen (§ 4·3) that for the method of paired comparisons,

$$\lambda = \tfrac{1}{4}t^3 \sum_i \delta_i^2 = \lim_{n\to\infty} \tfrac{1}{4}nt^3 \sum_i \delta_{in}^2/n,$$

and we can write, from (56),

$$\lambda'' = \lim_{n''\to\infty} \tfrac{1}{4}\pi n'' t^2(t-1) \sum_i \delta_{in''}^{*2}/n''.$$

The test situations are comparable when (57) holds and the asymptotic powers are identical when $\lambda = \lambda''$. Consequently the ratio of sample sizes for equal powers is asymptotically

$$\lim_{n,\, n''\to\infty} \frac{n''}{n} = \frac{t}{\pi(t-1)}, \tag{58}$$

and this is the asymptotic relative efficiency of the method of paired comparisons to the analysis of variance.

The theorems of Pitman and Noether yield relative efficiencies in terms of the efficacies of the tests being compared. For a statistic T_N based on a sample of size N, define

$$E(T_N) = \psi_N(\theta; T_N) \quad \text{and} \quad \operatorname{var}(T_N) = \sigma_N^2(\theta; T_N).$$

The efficacy of the test of H_0: $\theta = \theta_0$ against the alternative H: $\theta > \theta_0$ is

$$1/R_N^{ms}(\theta_0; T_N) = \{\psi^{(m)}(\theta_0; T_N)/\sigma_N(\theta_0; T_N)\}^{1/ms} \tag{59}$$

where $\theta_N = \theta_0 + k/N^s$. It is assumed that a particular alternative $\theta = \theta_N$ changes with the sample size N in such a way that $\lim_{N\to\infty} \theta_N = \theta_0$.

We have three tests of the hypothesis H_0: $\pi_i = 1/t$ with the alternative,

$$H_n: \quad \pi_i = t^{-1} + n^{-\frac{1}{2}}\delta_{in}, \quad \delta_{in} \to \delta_i \quad \text{as} \quad n \to \infty \quad (i = 1, \ldots, t).$$

The statistics, T_n for the method of paired comparisons, S_n for the multi-binomial procedure, and F_n for the analysis of variance, all have asymptotically non-central χ^2-distributions with parameters,

$$\lambda = \lim_{n\to\infty} \tfrac{1}{4}nt^3\theta_n, \quad \lambda' = \lim_{n\to\infty} \tfrac{1}{4}nt^3\theta_n, \quad \lambda'' = \lim_{n\to\infty} \tfrac{1}{4}n\pi t^2(t-1)\theta_n, \tag{60}$$

and degrees of freedom, $(t-1)$, $\frac{1}{2}t(t-1)$, $(t-1)$ respectively. In (60)

$$\theta_n = \sum_i \delta_{in}^2/n. \tag{61}$$

The subscripts were added to the statistics T, S and F in order to provide a notation consistent with the general definition (59) and to emphasize the dependence of the statistics on n.

The hypothesis H_0 is true when $\theta_n = \theta_0 = 0$. We may take $\psi_n(\theta; T_n)$, $\psi_n(\theta; S_n)$ and $\psi_n(\theta; F_n)$, corresponding to $\psi_N(\theta; T_N)$ in (59), respectively to be*

$$(t-1) + \tfrac{1}{4}nt^3\theta, \quad \tfrac{1}{2}t(t-1) + \tfrac{1}{4}nt^3\theta \quad \text{and} \quad (t-1) + \tfrac{1}{4}n\pi t^2(t-1)\theta.$$

Differentiation with respect to θ yields (for all values of θ including $\theta = 0$)

$$\psi_n'(0; T_n) = \tfrac{1}{4}nt^3, \quad \psi_n'(0; S_n) = \tfrac{1}{4}nt^3 \quad \text{and} \quad \psi_n'(0; F_n) = \tfrac{1}{4}n\pi t^2(t-1). \tag{62}$$

* Our values of ψ_n and σ_n^2, the latter given in (63), are not the means and variances of the statistics for finite n but rather were calculated from the limiting distributions. That this is permissible under conditions which may be verified for our tests was noted by Noether.

From the non-central χ^2-distributions,

$$\sigma_n^2(0; T_n) = 2(t-1), \quad \sigma_n^2(0; S_n) = t(t-1) \quad \text{and} \quad \sigma_n^2(0; F_n) = 2(t-1). \tag{63}$$

The efficacies of the three tests are, with $\delta = 1$, $m = 1$ in (59) in each case,

$$R_n(0; T_n) = \frac{nt^3}{\sqrt{[32(t-1)]}}, \quad R_n(0; S_n) = \frac{nt^2}{4}\sqrt{\frac{t}{t-1}} \quad \text{and} \quad R_n(0; F_n) = n\pi t^2\sqrt{\frac{t-1}{32}}. \tag{64}$$

The relative efficiencies of the three tests taken in pairs are given by the limits of the ratios of corresponding values of R_n subject to conditions set forth in the reference and which may be verified here. The relative efficiencies of S to T and to F are

$$\text{R.E. } (S \text{ to } T) = \lim_{n\to\infty} \frac{R_n(0; S_n)}{R_n(0; T_n)} = \left(\frac{2}{t}\right)^{\frac{1}{2}} \tag{65}$$

and

$$\text{R.E. } (S \text{ to } F) = \lim_{n\to\infty} \frac{R_n(0; S_n)}{R_n(0; F_n)} = \frac{(2t)^{\frac{1}{2}}}{\pi(t-1)}. \tag{66}$$

Table 1. *Asymptotic relative efficiencies*

(T, paired comparisons; S, multi-binomial; F, analysis of variance)

t	2	3	4	5	6	7	8	9	10	∞
T to F	63·7	47·7	42·4	39·8	38·2	37·1	36·4	35·8	35·4	31·8
S to T	100·0	81·7	70·7	63·2	57·7	53·5	50·0	47·1	44·7	0
S to F	63·7	39·0	29·9	25·1	22·0	19·7	18·1	16·9	15·9	0

The result obtained more directly for the relative efficiency of T to F may also be obtained using values given in (64). Asymptotic relative efficiencies expressed as percentages for values of t from 2 to 10 are given in Table 1.

When $t = 2$, the method of paired comparisons reduced to the statistical sign test and the relative efficiency (58) is $2/\pi$ as obtained by Dixon & Mood. The method of paired comparisons becomes more and more inefficient relative to the analysis of variance as t increases. This was to be expected in view of the work of Dixon & Mood for the sign test, and since the comparison was made for a situation in which the analysis of variance could appropriately be used. The assumptions of the analysis of variance are invalidated in much of the work on sensory difference and subjective testing, and it was to avoid those restrictive assumptions that the method of paired comparisons was developed. Further, although for $t = 2$, the tabled relative efficiency of T to F is only 63·7 %, Walsh (1946) has shown that, for samples of size $n = 4$, 5 and 6, the sign test is approximately 95 % as efficient as the 'Student' t-test, and, although the relative efficiency decreases as n increases, it is approximately 75 % when $n = 13$. Dixon & Mood (1946) and Dixon (1953) also indicate that the efficiency of the sign test is better for smaller sample sizes. It seems safe to assume that this situation also holds for $t > 2$ in paired comparisons.

The multi-binomial procedure is seen to have rapidly decreasing relative efficiencies as t increases in comparison with both the method of paired comparisons and the analysis of variance.

VI. Illustrative examples

6·1. *Estimated variances and covariances in a numerical example*

We shall use the data from the taste-testing experiment on the flavour characteristics of pork roasts as given as an example by Bradley & Terry (1952). We now assume that the data for the two judges may be pooled. The pork roasts were obtained from hogs fed corn rations and peanut supplements C, Cp and CP. The experiment, with the results of the two judges pooled, yielded the set of sums of ranks, 32, 28 and 30, respectively, for the rations while $n = 10$. The use of the tables in that paper showed the following values:

Σr_1	Σr_2	Σr_3	p_1	p_2	p_3	Prob.
32	28	30	0·24	0·43	0·32	0·63

This result is clearly non-significant at any realistic level of significance and is not indicative of any real ration effect on the flavours of the roasts.

We now obtain estimates of the variances and covariances of $\sqrt{n}\,p_1$, $\sqrt{n}\,p_2$ and $\sqrt{n}\,p_3$ or $\sqrt{n}\,(p_1-\pi_1)$, $\sqrt{n}\,(p_2-\pi_2)$ and $\sqrt{n}\,(p_3-\pi_3)$. The first computing step is to obtain values of $\hat{\lambda}_{ij}$ from substitution of values of p_i for π_i in (14). Then,

$$\begin{aligned} \hat{\lambda}_{11} &= 8{\cdot}243, & \hat{\lambda}_{22} &= 2{\cdot}566, \\ \hat{\lambda}_{12} &= -2{\cdot}228, & \hat{\lambda}_{23} &= -1{\cdot}778, \\ \hat{\lambda}_{13} &= -3{\cdot}189, & \hat{\lambda}_{33} &= 4{\cdot}780. \end{aligned}$$

The estimate of the determinant in the denominator of (23) is

$$\begin{vmatrix} 8{\cdot}243 & -2{\cdot}228 & -3{\cdot}189 & 1 \\ -2{\cdot}228 & 2{\cdot}566 & -1{\cdot}778 & 1 \\ -3{\cdot}189 & -1{\cdot}778 & 4{\cdot}780 & 1 \\ 1 & 1 & 1 & 0 \end{vmatrix} = -154{\cdot}976.$$

Now, for example, from (23),

$$\hat{\sigma}_{12} = \frac{1}{154{\cdot}976} \begin{vmatrix} -2{\cdot}228 & -1{\cdot}778 & 1 \\ -3{\cdot}189 & 4{\cdot}780 & 1 \\ 1 & 1 & 0 \end{vmatrix} = -0{\cdot}0485,$$

and similarly, the complete set of estimated variances and covariances is

$$\begin{aligned} \hat{\sigma}_{11} &= 0{\cdot}0703, & \hat{\sigma}_{22} &= 0{\cdot}1252, \\ \hat{\sigma}_{12} &= -0{\cdot}0485, & \hat{\sigma}_{23} &= -0{\cdot}0767, \\ \hat{\sigma}_{13} &= -0{\cdot}0218, & \hat{\sigma}_{33} &= 0{\cdot}0985. \end{aligned}$$

The estimated variances and covariances of p_1, p_2 and p_3 may be obtained by dividing those immediately above by $n = 10$. Consequently, the estimated standard errors of p_1, p_2 and p_3 are respectively given approximately by 0·084, 0·112 and 0·099.

A check on the computing may be made by computing the variance of $\sqrt{n}\sum_i p_i$ in terms of the variances and covariances of the elements of this sum. The result of course should be zero.

6·2. *Use of the power function*

(i) Suppose that the true values of the parameters π_1, π_2 and π_3 in the example of the preceding section are respectively 0·28, 0·59 and 0·13. (If these were estimates with $n = 10$, the significance level of the test would be 0·04.) We use δ_{in} defined in (34) as an approximation to δ_i and from (45) and (44) we have the approximate values

$$\lambda = 7{\cdot}43 \quad \text{and} \quad \phi = 1{\cdot}57.$$

The approximate power of a 0·05-level test, obtained from Tang's tables* with $f_1 = 2$, $f_2 = \infty$, $\alpha = 0{\cdot}05$, is $\beta(7{\cdot}43 \mid 0{\cdot}05, 2, \infty) = 0{\cdot}67$ and with $\alpha = 0{\cdot}01$, $\beta(7{\cdot}43 \mid 0{\cdot}01, 2, \infty) = 0{\cdot}56$ in the notation of (43). If we are interested in the probability of failing to reject H_0 when H_1 is true, we require $1 - \beta$. For $\alpha = 0{\cdot}05$, the probability of a type II error is 0·33 and, for $\alpha = 0{\cdot}01$, the probability of a type II error is 0·44.

(ii) Suppose that the differences observed in the estimates for the pork roast experiment are of practical importance (as distinct from statistical significance) and that for a 0·05-level test we desire the sample size for a second follow-up experiment such that $\beta = 0{\cdot}95$ or such that the probability of a type II error will be about 0·05. Without specifying n, we take the values of π_1, π_2 and π_3 to be 0·24, 0·43 and 0·32 as estimated in the experiment. Then from (45), $\lambda = 0{\cdot}123n$ and $\phi = \sqrt{(0{\cdot}123n/3)}$.

To determine n, we enter Tang's tables with $f_1 = 2, f_2 = \infty, \alpha = 0{\cdot}05$ and find that ϕ must have the value 2·35 for $\beta(0{\cdot}123n \mid 0{\cdot}05, 2, \infty) = 0{\cdot}95$. The number of repetitions required is obtained by setting

$$0{\cdot}123n/3 = (2{\cdot}35)^2,$$

from whence

$$n = 135.$$

This is of course a very large number of repetitions and is required owing to the small differences among the estimates of the experiment.

If we take the hypothetical case considered in (i) above and the same test requirements,

$$0{\cdot}743n/3 = (2{\cdot}35)^2 \quad \text{and} \quad n = 22.$$

This value may again seem high but it must be emphasized that most experiments are conducted without regard to the powers of the tests employed and this is true even when the analysis of variance is used.

VII. DISCUSSION AND SUMMARY

7·1. *Additional test procedures*

Two test procedures relating to paired comparisons were discussed by Bradley & Terry (1952) and have not yet been considered in this paper.

The first test is a combined test on the equality of treatment effects with the specification of the alternative hypothesis such that the parameters may differ from one group of repetitions to another. Suppose the paired comparisons experiment is performed in g groups of repetitions with n_u repetitions in the uth group, $\sum_{u=1}^{g} n_u = n$, and suppose that the treatments have parameters, $\pi_{1u}, \ldots, \pi_{tu}$, $\sum_i \pi_{iu} = 1$, in the uth group. The hypotheses are

$$H_0: \quad \pi_{iu} = 1/t \quad (i = 1, \ldots, t;\ u = 1, \ldots, g)$$

* Tang (1938) tabulated P_{II}, the probability of a type II error. In using the tables, linear interpolation was judged to be satisfactory for our examples.

and H_a: No π_{iu} assumed equal to any other π_{jv} ($i \neq j$ or $u \neq v$; $i, j = 1, \ldots, t$; $u, v = 1, \ldots, g$). Let T_u be the statistic corresponding to T for the uth group. Then

$$T_c = \sum_{u=1}^{g} T_u$$

is the appropriate statistic for this test. From the additive property of χ^2, it follows that under H_0, T_c has, for large n and fixed ratios n_u/n, the central χ^2-distribution with $g(t-1)$ degrees of freedom and, under H_a, T_c has the non-central χ^2-distribution with $g(t-1)$ degrees of freedom and parameter of non-centrality approximated by

$$\lambda_c = \tfrac{1}{4}t^3 \sum_{i} \sum_{u=1}^{g} n_u \left(\pi_{iu} - \frac{1}{t}\right)^2.$$

The second test is a test of agreement of ranking from group to group. The hypotheses are H_0: $\pi_{iu} = \pi_i$ ($i = 1, \ldots, t$; $u = 1, \ldots, g$) and H_a as specified in the preceding paragraph. Let T_p be the T-statistic obtained by pooling all groups of repetitions as though we have a homogeneous set of n repetitions. The appropriate statistic for this second test is $(T_c - T_p)$. This statistic has the central χ^2-distribution with $(g-1)(t-1)$ degrees of freedom under H_0. Under H_a, we state without proof that the distribution of $(T_c - T_p)$ is the non-central χ^2-distribution with $(g-1)(t-1)$ degrees of freedom and parameter of non-centrality approximated by

$$\lambda_u = \tfrac{1}{4}t^3 \sum_{i} \sum_{u=1}^{g} n_u(\pi_{iu}^2 - \pi_i^2).$$

7·2. *Discussion*

The results obtained in this paper are valid only for large samples. For small samples they may be approximately correct, but the difficulty of obtaining exact powers makes a check very nearly impossible. Under H_0 some information is available on the approach of the distribution of T to the central χ^2-distribution. The means and variances of T were computed from the exact small-sample distribution of T (Bradley, 1954*b*) and their approach to the corresponding moments of χ^2 noted. It appeared that the approximate tests would be adequate at least for $n \geqslant 10$.

In extreme cases, the estimators $p_1, \ldots, p_t$ sometimes have the set of values $1, 0, \ldots, 0$. This prohibits the computation of $\hat{\lambda}_{ij}$ and the estimation of variances and covariances of the estimators. This situation was previously discussed (Bradley, 1954*b*), and it was noted that the first treatment could be eliminated from the analysis and that secondary parameter estimates could be obtained for $\pi_2, \ldots, \pi_t$. This procedure again seems suitable when the specified set of values occurs. Then it will be possible to proceed with the analysis and the estimation of variances and covariances for the remaining estimates. Other extreme sets of sums of ranks yield extreme sets of estimators and methods of dealing with them are discussed in the cited reference.

The comparisons of the properties of the method of paired comparisons and of other test methods produce results that seem to be about what one would expect. The method of paired comparisons is clearly superior to the multi-binomial procedure both on the basis of the asymptotic powers of the tests for equal sample sizes and the study of their relative efficiency. It appears that the superiority of paired comparisons will become more marked as the number of treatments is increased. The comparison of the method of paired comparisons with the analysis of variance is less favourable to the method of paired comparisons.

Here the comparison was made when the analysis of variance would be the appropriate method. It should be remembered that the method of paired comparisons was formulated largely for subjective tests and it is there that the assumptions of the analysis of variance are seriously suspect. Further, even if a scoring technique could be devised for which the analysis of variance could be used, it is quite possible that an over-all consideration of time of judging and analysis would dictate the use of more samples and the non-parametric method. The large-sample results indicate that for the usual numbers of treatments $2\frac{1}{2}$ times as many samples are required for the method of paired comparisons as for the analysis of variance. It is conceivable in a tasting experiment, for example, that an individual can indicate rankings in a paired-treatment experiment as contrasted to scoring individual samples on an 11-point scale, say, with much less taste fatigue and with a speed ratio in favour of ranking in excess of the ratio of sample numbers.

7·3. *Summary*

In this paper we have examined some of the properties of the method of paired comparisons. The results obtained are asymptotically correct for large-sample sizes or for large numbers of repetitions of the paired-comparisons design. Formulae for the variances and covariances of estimates of treatment ratings, $\pi_1, \ldots, \pi_t$, have been obtained, and these were not heretofore available. It has been shown that statistics previously used for tests of significance have limiting non-central χ^2-distributions, and the appropriate parameters of non-centrality are given.

It was found that in comparison with the analysis of variance the relative efficiency of the method of paired comparisons is $t/\{\pi(t-1)\}$, and, when $t = 2$, this has the value $2/\pi$ previously obtained in comparing the sign test with the 'Student' t-test. The method of paired comparisons was seen to be considerably better than a multi-binomial procedure postulated and the asymptotic powers of these two tests are plotted for a number of examples along with similar values for the analysis of variance.

Examples of the use of the power function developed were given in application to an experiment in taste testing. Estimated variances and covariances of the estimators of the example were computed.

REFERENCES

ABELSON, R. M. & BRADLEY, R. A. (1954). A 2×2 factorial with paired comparisons. *Biometrics*, **10**, 487.

BRADLEY, R. A. (1953). Some statistical methods in taste testing and quality evaluation. *Biometrics*, **9**, 22.

BRADLEY, R. A. (1954*a*). Incomplete block rank analysis: on the appropriateness of the model for a method of paired comparisons. *Biometrics*, **10**, 375.

BRADLEY, R. A. (1954*b*). Rank analysis of incomplete block designs. II. Additional tables for the method of paired comparisons. *Biometrika*, **41**, 502.

BRADLEY, R. A. & TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, **39**, 324.

CHANDA, K. C. (1954). A note on the consistency and maxima of the roots of likelihood equations. *Biometrika*, **41**, 56.

COCHRAN, W. G. (1937). The efficiencies of the binomial series test of significance of a mean and of a correlation coefficient. *J. R. Statist. Soc.* A, **100**, 69.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

DIXON, W. J. (1953). Power functions of the sign test and power efficiency for normal alternatives. *Ann. Math. Statist.* **24**, 467.

DIXON, W. J. & MOOD, A. M. (1946). The statistical sign test. *J. Amer. Statist. Ass.* **41**, 557.

FIX, EVELYN (1949). Tables of noncentral χ^2. *Univ. Calif. Publ. Statist.* **1** (2), 15.

HALD, A. (1952). *Statistical Theory with Engineering Applications*, §9·9. New York: John Wiley and Sons, Inc.

HEMELRIJK, J. (1952). A theorem on the sign test when ties are present. *Indag. Math.* **14**, 322.

HOPKINS, J. W. (1954). Incomplete block rank analysis: some taste test results. *Biometrics*, **10**, 391.

MOSTELLER, F. (1951). Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, **16**, 3.

NOETHER, G. (1955). On a theorem of Pitman. *Ann. Math. Statist.* **26**, 64.

PEARSON, E. S. & HARTLEY, H. O. (1951). Charts for the power function for analysis of variance tests, revised from the non-central *F*-distribution. *Biometrika*, **38**, 112.

PITMAN, E. J. G. (1948). *Lecture Notes on Non-Parametric Statistics*. Columbia University, New York.

TANG, P. C. (1938). The power function of analysis of variance tests with tables and illustrations of their use. *Statist. Res. Mem.* **2**, 126.

TERRY, M. E., BRADLEY, R. A. & DAVIS, L. L. (1952). New designs and techniques for organoleptic testing. *Food Tech., Lond.*, **6**, 250.

WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54**, 426.

WALSH, J. (1946). On the power function of the sign test for slippage of means. *Ann. Math Statist.* **17**, 358.

A METHOD OF ASSIGNING CONFIDENCE LIMITS TO LINEAR COMBINATIONS OF VARIANCES

BY A. HUITSON, PH.D.

Royal Naval Scientific Service

1. INTRODUCTION AND SUMMARY

Linear combinations of variances occur frequently in analysis of variance work, when we are considering the estimation either of individual components of variance or of gross variability. Percentage points for such functions cannot be computed directly in the same manner as that adopted when the χ^2 distribution is used to give percentage points in problems where only a single unknown variance is involved.

It is the purpose of the present paper (*a*) to derive a series approximation to the percentage points of functions of this type, (*b*) to investigate the numerical behaviour of the expansion in some particular cases and (*c*) to provide tables of the series expansion, suitable for use in certain practical problems. An example is given in which use is made of the tables, together with some general discussion of the single-factor type of experiment.

2. DEVELOPMENT OF A SERIES APPROXIMATION

Let us consider any number of unknown population variances which we shall denote by $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$.*

Suppose that the observed data provide estimates s_i^2 of these variances based on f_i degrees of freedom respectively, and that these estimates are distributed independently of each other in the form

$$p(s_i^2)\,ds_i^2 = \frac{1}{\Gamma(\frac{1}{2}f_i)}\left(\frac{f_i s_i^2}{2\sigma_i^2}\right)^{\frac{1}{2}f_i - 1} \exp\left[-\frac{1}{2}\frac{f_i s_i^2}{\sigma_i^2}\right] d\left(\frac{f_i s_i^2}{2\sigma_i^2}\right). \tag{1}$$

We are seeking to assign confidence limits, calculable from the observations, to the expression $\sum_{i=1}^{r} \lambda_i \sigma_i^2$, where $\lambda_1 \ldots \lambda_r$ are known arbitrary constants. In order to accomplish this, we shall seek a function y, which will satisfy the equation

$$P_r\left[\frac{\sum_{i=1}^{r} \lambda_i s_i^2}{\sum_{i=1}^{r} \lambda_i \sigma_i^2} < y(s_1^2, s_2^2, \ldots, s_r^2, P)\right] = P, \tag{2}$$

where P_r is used to denote 'the probability of the relation in the bracket following'. It is by no means obvious that there is a function y having the required property, but as we shall see later it is certainly possible to find various functions which have *approximately* the required properties. We shall be content to leave open the question as to whether our problem may be solved exactly.

* More strictly, in most practical applications, the $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$ are parameters which are already linear functions of more basic variances. It will be convenient, however, to refer to the $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$ themselves as variances since they have estimates $s_1^2, s_2^2, \ldots, s_r^2$ which are distributed in the form (1).

The large-sample normal approximation to y is

$$1+\xi\frac{\sqrt{2}}{(\Sigma\lambda_i\sigma_i^2)}\sqrt{\left(\Sigma\lambda_i^2\frac{\sigma_i^4}{f_i}\right)},$$

where ξ represents the Pth percentage point of the unit normal deviate. However, as the variances are unknown, it will be convenient to take as our initial approximation to y

$$1+\xi\frac{\sqrt{2}}{(\Sigma\lambda_i s_i^2)}\sqrt{\left(\Sigma\lambda_i^2\frac{s_i^4}{f_i}\right)}. \tag{3}$$

All summations are to be taken over the range 1 to r.

We shall now introduce the notation

$$V_{mn}=\frac{\Sigma\dfrac{\lambda_i^m s_i^{2m}}{f_i^n}}{(\Sigma\lambda_i s_i^2)^m}=\Sigma\frac{c_i^m}{f_i^n}, \quad \text{where} \quad c_i=\frac{\lambda_i s_i^2}{(\Sigma\lambda_i s_i^2)},$$

and denote the corresponding functions of the unknown variances by Δ_{mn} and γ_i respectively. Thus our initial approximation to y becomes

$$1+\xi\sqrt{(2V_{21})}. \tag{4}$$

If we assume that y can be written in the form

$$y=h_0+h_1+h_2+h_3+\ldots, \tag{5}$$

where h_i is of order $f^{-\frac{1}{2}i}$, then the large-sample normal approximation implies that

$$\left.\begin{aligned} h_0&=1,\\ h_1&=\xi\sqrt{(2V_{21})}.\end{aligned}\right\} \tag{6}$$

We shall now proceed to derive the next term, h_2, in the series expansion of y. Let us consider the new variable

$$u=\frac{\Sigma\lambda_i s_i^2}{\Sigma\lambda_i\sigma_i^2}-(h_0+h_1). \tag{7}$$

As u is a function of the variance estimates, the moment-generating function of u may be found by averaging over the distributions of the s_i^2. On simplification we find the first three cumulants of u to be

$$K_1=-\xi\sqrt{(2\Delta_{21})} \quad \text{to order } f^{-1},$$

$$K_2=2\Delta_{21}+4\sqrt{2}\,\xi\{(\Delta_{21})^{\frac{3}{2}}-\Delta_{32}(\Delta_{21})^{-\frac{1}{2}}\} \quad \text{to order } f^{-\frac{3}{2}},$$

and

$$K_3=8\Delta_{32} \quad \text{to order } f^{-2}.$$

Thus we can see that $K_1=O(f^{-\frac{1}{2}})$, $K_2=O(f^{-1})$ and $K_3=O(f^{-2})$. It appears that u has cumulants which decrease in powers of f such that $K_r=O(f^{-r+1})$.

Using the Cornish & Fisher expansion (1937) we then obtain after simplification the Pth percentage point of u, u_P say, to be equal to

$$\xi^2\left[2\Delta_{21}-\frac{4}{3}\frac{\Delta_{32}}{\Delta_{21}}\right]-\frac{2}{3}\frac{\Delta_{32}}{\Delta_{21}} \quad \text{to order } \left(\frac{1}{f}\right). \tag{8}$$

A METHOD OF ASSIGNING CONFIDENCE LIMITS TO LINEAR COMBINATIONS OF VARIANCES

BY A. HUITSON, PH.D.

Royal Naval Scientific Service

1. INTRODUCTION AND SUMMARY

Linear combinations of variances occur frequently in analysis of variance work, when we are considering the estimation either of individual components of variance or of gross variability. Percentage points for such functions cannot be computed directly in the same manner as that adopted when the χ^2 distribution is used to give percentage points in problems where only a single unknown variance is involved.

It is the purpose of the present paper (*a*) to derive a series approximation to the percentage points of functions of this type, (*b*) to investigate the numerical behaviour of the expansion in some particular cases and (*c*) to provide tables of the series expansion, suitable for use in certain practical problems. An example is given in which use is made of the tables, together with some general discussion of the single-factor type of experiment.

2. DEVELOPMENT OF A SERIES APPROXIMATION

Let us consider any number of unknown population variances which we shall denote by $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$.*

Suppose that the observed data provide estimates s_i^2 of these variances based on f_i degrees of freedom respectively, and that these estimates are distributed independently of each other in the form

$$p(s_i^2)\,ds_i^2 = \frac{1}{\Gamma(\frac{1}{2}f_i)}\left(\frac{f_i s_i^2}{2\sigma_i^2}\right)^{\frac{1}{2}f_i-1}\exp\left[-\frac{1}{2}\frac{f_i s_i^2}{\sigma_i^2}\right]d\left(\frac{f_i s_i^2}{2\sigma_i^2}\right). \tag{1}$$

We are seeking to assign confidence limits, calculable from the observations, to the expression $\sum_{i=1}^{r}\lambda_i\sigma_i^2$, where $\lambda_1 \ldots \lambda_r$ are known arbitrary constants. In order to accomplish this, we shall seek a function y, which will satisfy the equation

$$P_r\left[\frac{\sum_{i=1}^{r}\lambda_i s_i^2}{\sum_{i=1}^{r}\lambda_i\sigma_i^2} < y(s_1^2, s_2^2, \ldots, s_r^2, P)\right] = P, \tag{2}$$

where P_r is used to denote 'the probability of the relation in the bracket following'. It is by no means obvious that there is a function y having the required property, but as we shall see later it is certainly possible to find various functions which have *approximately* the required properties. We shall be content to leave open the question as to whether our problem may be solved exactly.

* More strictly, in most practical applications, the $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$ are parameters which are already linear functions of more basic variances. It will be convenient, however, to refer to the $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$ themselves as variances since they have estimates $s_1^2, s_2^2, \ldots, s_r^2$ which are distributed in the form (1).

The large-sample normal approximation to y is

$$1+\xi\frac{\sqrt{2}}{(\Sigma\lambda_i\sigma_i^2)}\sqrt{\left(\Sigma\lambda_i^2\frac{\sigma_i^4}{f_i}\right)},$$

where ξ represents the Pth percentage point of the unit normal deviate. However, as the variances are unknown, it will be convenient to take as our initial approximation to y

$$1+\xi\frac{\sqrt{2}}{(\Sigma\lambda_i s_i^2)}\sqrt{\left(\Sigma\lambda_i^2\frac{s_i^4}{f_i}\right)}. \tag{3}$$

All summations are to be taken over the range 1 to r.

We shall now introduce the notation

$$V_{mn}=\frac{\Sigma\dfrac{\lambda_i^m s_i^{2m}}{f_i^n}}{(\Sigma\lambda_i s_i^2)^m}=\Sigma\frac{c_i^m}{f_i^n},\quad\text{where}\quad c_i=\frac{\lambda_i s_i^2}{(\Sigma\lambda_i s_i^2)},$$

and denote the corresponding functions of the unknown variances by Δ_{mn} and γ_i respectively. Thus our initial approximation to y becomes

$$1+\xi\sqrt{(2V_{21})}. \tag{4}$$

If we assume that y can be written in the form

$$y=h_0+h_1+h_2+h_3+\ldots, \tag{5}$$

where h_i is of order $f^{-\frac{1}{2}i}$, then the large-sample normal approximation implies that

$$\left.\begin{aligned}h_0&=1,\\ h_1&=\xi\sqrt{(2V_{21})}.\end{aligned}\right\} \tag{6}$$

We shall now proceed to derive the next term, h_2, in the series expansion of y. Let us consider the new variable

$$u=\frac{\Sigma\lambda_i s_i^2}{\Sigma\lambda_i\sigma_i^2}-(h_0+h_1). \tag{7}$$

As u is a function of the variance estimates, the moment-generating function of u may be found by averaging over the distributions of the s_i^2. On simplification we find the first three cumulants of u to be

$$K_1=-\xi\sqrt{(2\Delta_{21})}\quad\text{to order } f^{-1},$$

$$K_2=2\Delta_{21}+4\sqrt{2}\,\xi\{(\Delta_{21})^{\frac{3}{2}}-\Delta_{32}(\Delta_{21})^{-\frac{1}{2}}\}\quad\text{to order } f^{-\frac{3}{2}},$$

and

$$K_3=8\Delta_{32}\quad\text{to order } f^{-2}.$$

Thus we can see that $K_1=O(f^{-\frac{1}{2}})$, $K_2=O(f^{-1})$ and $K_3=O(f^{-2})$. It appears that u has cumulants which decrease in powers of f such that $K_r=O(f^{-r+1})$.

Using the Cornish & Fisher expansion (1937) we then obtain after simplification the Pth percentage point of u, u_P say, to be equal to

$$\xi^2\left[2\Delta_{21}-\frac{4}{3}\frac{\Delta_{32}}{\Delta_{21}}\right]-\frac{2}{3}\frac{\Delta_{32}}{\Delta_{21}}\quad\text{to order }\left(\frac{1}{f}\right). \tag{8}$$

But this is a function of the ratios of the unknown variances and thus is not of immediate use to us. Consider any particular ratio, say σ_m^2/σ_n^2, where σ_m^2 and σ_n^2 are any two of the $\sigma_1^2 \dots \sigma_r^2$. The mean of s_m^2/s_n^2 is equal to

$$\frac{\sigma_m^2}{\sigma_n^2}\left(\frac{f_n}{f_n-2}\right) = \frac{\sigma_m^2}{\sigma_n^2}\left(1+\frac{2}{f_n}+\dots\right).$$

Also the variance of s_m^2/s_n^2 is of the order of f^{-1}. Therefore neglecting terms of order f^{-1} and lower we have in the probability sense $s_m^2/s_n^2 = \sigma_m^2/\sigma_n^2$. Thus, without affecting the accuracy of equation (8) we can replace all the ratios of the unknown variances by the ratios of their estimates, since all terms on the right-hand side of (8) are of the same order.

Therefore Δ_{mn} will now become V_{mn} and

$$h_2 = \xi^2\left[2V_{21} - \frac{4}{3}\frac{V_{32}}{V_{21}}\right] - \frac{2}{3}\frac{V_{32}}{V_{21}}. \tag{9}$$

The fact that (9) is of order f^{-1} and not of higher order is a verification that we have taken the correct value for h_1. Thus the method may be regarded as self-checking, in that, in each application, it checks the value which has been obtained previously and adds another term to the series. Using the same method the next term of the series expansion for y, h_3, has been found to be

$$\begin{aligned} h_3 = \sqrt{(2V_{21})}\{&\xi[-\tfrac{4}{3}V_{32}(V_{21})^{-1} + \tfrac{9}{2}V_{43}(V_{21})^{-2} - V_{22}(V_{21})^{-1} - \tfrac{2}{9}\tfrac{3}{9}(V_{32})^2(V_{21})^{-3}] \\ &+ \xi^3[-\tfrac{8}{3}V_{32}(V_{21})^{-1} + 2V_{21} - \tfrac{1}{9}\tfrac{6}{9}V_{32}^2(V_{21})^{-3} + \tfrac{5}{2}V_{43}(V_{21})^{-2}]\}. \end{aligned} \tag{10}$$

However, the solution was carried no further as the work involved would have been too laborious.

3. CHECKS

The following algebraic checks have been carried out on the series expansion of y derived above.

(*a*) It has been shown to agree with an independent development of the series solution carried through by Welch (1956) for the same problem, although his treatment and the algebra involved are somewhat different in detail from mine.

(*b*) It has also been shown to agree with a result published by Bartlett (1953). In this paper Bartlett derives a confidence interval for a function of the form $\sigma_1^2 + \lambda\sigma_2^2$. To the appropriate order of terms, $h_0 + h_1 + h_2$ is the solution of this equation.

(*c*) A simple check, which has also been carried out, is to consider the special case when all the degrees of freedom except one are infinite. A series expansion for the function $\Sigma\lambda_i\sigma_i^2$ can then be found from the χ^2 distribution and this compared term by term with the series derived above.

4. NUMERICAL TESTING

The series approximation given by $h_0 + h_1 + h_2 + h_3$ is applicable to the general problem involving r variances. In order to carry out a numerical investigation of it, we shall limit ourselves to the particular case $r = 2$ only and we shall assume both λ's have the same sign.

We can now write

$$V_{mn} = \frac{c^m}{f_1^n} + \frac{(1-c)^m}{f_2^n}, \quad \text{where} \quad c = \frac{\lambda_1 s_1^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2} \quad \text{and} \quad \gamma = \frac{\lambda_1 \sigma_1^2}{\lambda_1 \sigma_1^2 + \lambda_2 \sigma_2^2}.$$

For the cases in which both the λ's are of the same sign, the permissible values of c lie between 0 and 1.

Consider an approximation to the Pth percentage point of $(\lambda_1 s_1^2+\lambda_2 s_2^2)/(\lambda_1\sigma_1^2+\lambda_2\sigma_2^2)$ depending only on the sample statistic c (it will also depend on the probability P), $g(c)$ say. To test the accuracy of this approximation, we must evaluate

$$P_r\left[\frac{\lambda_1 s_1^2+\lambda_2 s_2^2}{\lambda_1\sigma_1^2+\lambda_2\sigma_2^2} < g(c)\right] = Q \quad \text{say}, \tag{11}$$

in order to see how close it is to P. Strictly, such a function as $g(c)$ is not an approximation to the Pth percentage point of $(\lambda_1 s_1^2+\lambda_2 s_2^2)/(\lambda_1\sigma_1^2+\lambda_2\sigma_2^2)$, as this is a function of the unknown variances while $g(c)$ is independent of these variances. It can only be regarded as such an approximation in the sense that it makes the value of Q in equation (11) close to P.

It can be shown by straightforward transformations that

$$Q = \int_0^1 I_{f_1+f_2}\left[\left(\frac{\gamma b}{f_1}+\frac{(1-\gamma)(1-b)}{f_2}\right)^{-1} g\left\{\frac{f_2\gamma b}{f_2\gamma b+f_1(1-\gamma)(1-b)}\right\}\right] p(b)\,db, \tag{12}$$

where

$$I_F[x] = \int_0^x \frac{1}{\Gamma(\frac{1}{2}F)}(\tfrac{1}{2}y)^{\frac{1}{2}F-1}e^{-\frac{1}{2}y}d(\tfrac{1}{2}y),$$

$$p(b)\,db = \frac{1}{B(\frac{1}{2}f_1,\frac{1}{2}f_2)}b^{\frac{1}{2}f_1-1}(1-b)^{\frac{1}{2}f_2-1}\,db$$

and

$$b = \frac{f_1 s_1^2/\sigma_1^2}{f_1 s_1^2/\sigma_1^2+f_2 s_2^2/\sigma_2^2}.$$

From equation (12) the following values of Q were obtained by quadrature when $f_1 = f_2 = 10$ with $g(c)$ replaced by the series approximation, cut off at various points, in order to see what effect each successive term of the series had on the result:

Values of Q when $f_1 = f_2 = 10$

	$P=0{\cdot}05$		$P=0{\cdot}95$	
	$\gamma=0{\cdot}5$	$\gamma=0{\cdot}2$	$\gamma=0{\cdot}5$	$\gamma=0{\cdot}2$
$g(c)=h_0+h_1$	0·0197	0·0269	0·9424	0·9419
$g(c)=h_0+h_1+h_2$	0·0358	0·0469	0·9547	0·9533
$g(c)=h_0+h_1+h_2+h_3$	0·0453	0·0544	0·9497	0·9501

The values of Q for $\gamma=\alpha$, say, and $\gamma=1-\alpha$ will be identical when $f_1=f_2$. Therefore the values of Q when $\gamma=0{\cdot}8$ will correspond to those given in the third and last columns of the above table. In all except the third column it is seen that the addition of each successive term to the series approximation brings the value of Q closer to the chosen value of P. In the third column the addition of h_3 has moved Q in the right direction, but too far.

It may also be noted that the values of Q obtained when $P=0{\cdot}05$ are much more in error than the corresponding ones for $P=0{\cdot}95$. This suggests that the series approximation may not be so good at the lower end of the distribution as it is at the upper end.

This comparison relates only to the case $r = 2$, and λ_1 and λ_2 positive. More work would be necessary to assess the merits of the series for the more general situation. The results for the case considered are, however, sufficiently promising to make it appear worth while to table some results for the larger numbers of degrees of freedom at least. Such tables are presented in the next section.

5. TABLES

A two-decimal table will usually be quite adequate for a working statistician. Four probability levels are tabled in this paper (Tables 1–4).

We shall regard the series solution as giving us two-decimal accuracy when $h_3 < 0{\cdot}005$. The argument here is that if h_3 is negligible, then so, almost certainly, are higher terms of the series. If h_3 is not negligible, it does not follow that the higher terms are not negligible, but we have no absolute justification for assuming that they are. Using this criterion, the tabled values will be correct, except possibly those for which one of the degrees of freedom is 16, which may be one or two units in error in the second decimal place.

The values of degrees of freedom for which the tables have been constructed are such that their square roots form a harmonic set. The purpose of this is to make interpolation with respect to the degrees of freedom easier. Thus, for intermediate values of f, it is necessary to interpolate with respect to $12/\sqrt{f}$. Tables 1–4 can be used to assign 90 and 98 % confidence intervals to any sum of two variances, provided the degrees of freedom of the variance estimates are large anough.

6. EXAMPLE

The following example is given by Tippett (1952, p. 111). It is of the single-factor analysis type. Mule cops (bobbins) of cotton yarn were collected in blocks of 20, each block being from a different mule. Two leas (1 lea = 120 yards) were weighed from each cop, giving for each block 39 total degrees of freedom, 19 'between cops' and 20 'within cops', together with corresponding sums of squares. There were six such blocks and the six sets of sums of squares and degrees of freedom were added to give the following table:

Source of variation	Sum of squares	Degrees of freedom	Mean square	Expected value of mean square
Between cops (within blocks)	19,138·85	114	$167{\cdot}88 = s_1^2$	$\sigma_1^2 = 2\sigma_E^2 + \sigma_I^2$
Within cops	5,681·00	120	$47{\cdot}34 = s_2^2$	$\sigma_2^2 = \sigma_I^2$
Total	24,819·85	234		

where σ_E^2 is the variance between cops and σ_I^2 is the interaction (or error) variance in the usual manner.

The mean squares give an analysis of the variations within the blocks, and these are interesting because they are due to factors not easily controllable. The block to block variations are not studied because they can be easily eliminated by careful adjustment of the mules.

Table 1. *Lower 5 % critical values of* $(\lambda_1 s_1^2 + \lambda_2 s_2^2)/(\lambda_1 \sigma_1^2 + \lambda_2 \sigma_2^2)$

$\frac{\lambda_1 s_1^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2}$		0·0	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1·0
f_1	f_2											
16	16	0·50	0·52	0·55	0·59	0·62	0·64	0·62	0·59	0·55	0·52	0·50
	36	0·65	0·67	0·69	0·71	0·70	0·66	0·62	0·58	0·55	0·52	0·50
	144	0·81	0·82	0·81	0·76	0·71	0·66	0·62	0·58	0·55	0·52	0·50
	∞	1·00	0·91	0·83	0·77	0·71	0·66	0·62	0·58	0·55	0·52	0·50
36	16	0·50	0·52	0·55	0·58	0·62	0·66	0·70	0·71	0·69	0·67	0·65
	36	0·65	0·67	0·69	0·72	0·74	0·75	0·74	0·72	0·69	0·67	0·65
	144	0·81	0·83	0·83	0·83	0·81	0·78	0·75	0·72	0·69	0·67	0·65
	∞	1·00	0·95	0·90	0·86	0·82	0·78	0·75	0·72	0·70	0·67	0·65
144	16	0·50	0·52	0·55	0·58	0·62	0·66	0·71	0·76	0·81	0·82	0·81
	36	0·65	0·67	0·69	0·72	0·75	0·78	0·81	0·83	0·83	0·83	0·81
	144	0·81	0·83	0·84	0·86	0·86	0·87	0·86	0·86	0·84	0·83	0·81
	∞	1·00	0·98	0·96	0·94	0·92	0·90	0·88	0·86	0·85	0·83	0·81
∞	16	0·50	0·52	0·55	0·58	0·62	0·66	0·71	0·77	0·83	0·91	1·00
	36	0·65	0·67	0·70	0·72	0·75	0·78	0·82	0·86	0·90	0·95	1·00
	144	0·81	0·83	0·85	0·86	0·88	0·90	0·92	0·94	0·96	0·98	1·00
	∞	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00

Table 2. *Upper 5 % critical values of* $(\lambda_1 s_1^2 + \lambda_2 s_2^2)/(\lambda_1 \sigma_1^2 + \lambda_2 \sigma_2^2)$

$\frac{\lambda_1 s_1^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2}$		0·0	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1·0
f_1	f_2											
16	16	1·64	1·56	1·50	1·46	1·44	1·43	1·44	1·46	1·50	1·56	1·64
	36	1·42	1·37	1·35	1·34	1·34	1·36	1·39	1·43	1·48	1·55	1·64
	144	1·20	1·19	1·20	1·22	1·25	1·29	1·33	1·39	1·46	1·55	1·64
	∞	1·00	1·04	1·09	1·13	1·18	1·24	1·30	1·37	1·45	1·54	1·64
36	16	1·64	1·55	1·48	1·43	1·39	1·36	1·34	1·34	1·35	1·37	1·42
	36	1·42	1·37	1·33	1·31	1·29	1·29	1·29	1·31	1·33	1·37	1·42
	144	1·20	1·18	1·18	1·18	1·20	1·21	1·24	1·27	1·31	1·36	1·42
	∞	1·00	1·03	1·06	1·10	1·13	1·17	1·21	1·26	1·31	1·36	1·42
144	16	1·64	1·55	1·46	1·39	1·33	1·29	1·25	1·22	1·20	1·19	1·20
	36	1·42	1·36	1·31	1·27	1·24	1·21	1·20	1·18	1·18	1·18	1·20
	144	1·20	1·18	1·16	1·15	1·14	1·14	1·14	1·15	1·16	1·18	1·20
	∞	1·00	1·02	1·03	1·05	1·07	1·09	1·11	1·13	1·15	1·18	1·20
∞	16	1·64	1·54	1·45	1·37	1·30	1·24	1·18	1·13	1·09	1·04	1·00
	36	1·42	1·36	1·31	1·26	1·21	1·17	1·13	1·10	1·06	1·03	1·00
	144	1·20	1·18	1·15	1·13	1·11	1·09	1·07	1·05	1·03	1·02	1·00
	∞	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00

Table 3. *Lower* 1 % *critical values of* $(\lambda_1 s_1^2 + \lambda_2 s_2^2)/(\lambda_1 \sigma_1^2 + \lambda_2 \sigma_2^2)$

$\frac{\lambda_1 s_1^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2}$		0·0	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1·0
f_1	f_2											
16	16	0·37	0·39	0·41	0·45	0·50	0·53	0·50	0·45	0·41	0·39	0·37
	36	0·53	0·56	0·59	0·61	0·59	0·54	0·48	0·44	0·41	0·39	0·37
	144	0·75	0·76	0·74	0·66	0·58	0·52	0·47	0·44	0·41	0·39	0·37
	∞	1·00	0·86	0·74	0·65	0·58	0·52	0·48	0·44	0·42	0·39	0·37
36	16	0·37	0·39	0·41	0·44	0·48	0·54	0·59	0·61	0·59	0·56	0·53
	36	0·53	0·56	0·59	0·62	0·65	0·66	0·65	0·62	0·59	0·56	0·53
	144	0·75	0·76	0·77	0·76	0·73	0·69	0·65	0·62	0·59	0·56	0·53
	∞	1·00	0·92	0·85	0·79	0·74	0·69	0·65	0·62	0·59	0·56	0·53
144	16	0·37	0·39	0·41	0·44	0·47	0·52	0·58	0·66	0·74	0·76	0·75
	36	0·53	0·56	0·59	0·62	0·65	0·69	0·73	0·76	0·77	0·76	0·75
	144	0·75	0·77	0·78	0·80	0·81	0·82	0·81	0·80	0·78	0·77	0·75
	∞	1·00	0·97	0·94	0·91	0·88	0·85	0·83	0·81	0·79	0·77	0·75
∞	16	0·37	0·39	0·42	0·44	0·48	0·52	0·58	0·65	0·74	0·86	1·00
	36	0·53	0·56	0·59	0·62	0·65	0·69	0·74	0·79	0·85	0·92	1·00
	144	0·75	0·77	0·79	0·81	0·83	0·85	0·88	0·91	0·94	0·97	1·00
	∞	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00

Table 4. *Upper* 1 % *critical values of* $(\lambda_1 s_1^2 + \lambda_2 s_2^2)/(\lambda_1 \sigma_1^2 + \lambda_2 \sigma_2^2)$

$\frac{\lambda_1 s_1^2}{\lambda_1 s_1^2 + \lambda_2 s_2^2}$		0·0	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1·0
f_1	f_2											
16	16	2·00	1·85	1·76	1·70	1·67	1·65	1·67	1·70	1·76	1·85	2·00
	36	1·63	1·55	1·52	1·50	1·51	1·54	1·58	1·63	1·71	1·83	2·00
	144	1·29	1·28	1·29	1·32	1·36	1·41	1·47	1·56	1·68	1·84	2·00
	∞	1·00	1·06	1·11	1·18	1·25	1·33	1·42	1·53	1·66	1·82	2·00
36	16	2·00	1·83	1·71	1·63	1·58	1·54	1·51	1·50	1·52	1·55	1·63
	36	1·63	1·54	1·49	1·45	1·43	1·42	1·43	1·45	1·49	1·54	1·63
	144	1·29	1·27	1·26	1·27	1·28	1·31	1·35	1·40	1·46	1·54	1·63
	∞	1·00	1·04	1·08	1·13	1·18	1·24	1·30	1·37	1·45	1·53	1·63
144	16	2·00	1·82	1·68	1·56	1·47	1·41	1·36	1·32	1·29	1·28	1·29
	36	1·63	1·54	1·46	1·40	1·35	1·31	1·28	1·27	1·26	1·27	1·29
	144	1·29	1·26	1·24	1·22	1·21	1·20	1·21	1·22	1·24	1·26	1·29
	∞	1·00	1·02	1·05	1·07	1·10	1·13	1·16	1·19	1·22	1·26	1·29
∞	16	2·00	1·82	1·66	1·53	1·42	1·33	1·25	1·18	1·11	1·06	1·00
	36	1·63	1·53	1·45	1·37	1·30	1·24	1·18	1·13	1·08	1·04	1·00
	144	1·29	1·26	1·22	1·19	1·16	1·13	1·10	1·07	1·05	1·02	1·00
	∞	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00	1·00

'Between cops' corresponds to the factor and 'within cops' to the error; but it is inappropriate to speak of error when the investigation is not a controlled, or partially controlled, experiment.

Suppose that we wish to find confidence limits for the gross variability (i.e. the variance of a single observation). The estimate of the gross variability, $\sigma_E^2+\sigma_I^2 = \frac{1}{2}\sigma_1^2+\frac{1}{2}\sigma_2^2$, is

$$\lambda_1 s_1^2+\lambda_2 s_2^2 = \tfrac{1}{2}(47{\cdot}34)+\tfrac{1}{2}(167{\cdot}88) = 107{\cdot}61.$$

Then
$$c = \lambda_1 s_1^2/(\lambda_1 s_1^2+\lambda_2 s_2^2) = 0{\cdot}2200, \quad f_1 = 120, \quad f_2 = 114.$$

Interpolating in Table 1 we get a value of 0·83, and from Table 2 we get 1·18. Hence

$$P_r\left\{0{\cdot}83 < \frac{\lambda_1 s_1^2+\lambda_2 s_2^2}{\sigma_1^2+\sigma_2^2} < 1{\cdot}18\right\} \doteq 0{\cdot}90,$$

or
$$P_r\left\{\frac{107{\cdot}61}{1{\cdot}18} < \sigma_1^2+\sigma_2^2 < \frac{107{\cdot}61}{0{\cdot}83}\right\} \doteq 0{\cdot}90,$$

or
$$P_r\{91 < \sigma_1^2+\sigma_2^2 < 130\} \doteq 0{\cdot}90.$$

Thus 90 % confidence limits for $\sqrt{(\sigma_1^2+\sigma_2^2)}$ are approximately 9·5 and 11·4.

In a similar manner 98 % confidence limits can be found by using Tables 3 and 4.

7. Discussion

In order to use the tables given in this paper in experiments of the type described above, it is necessary that both rows of the analysis of variance table should have degrees of freedom as large as 16. This can be ensured by making the number of levels of the factor large enough. If the number of levels is small we shall not get a good estimate of the variability between levels, and hence we shall be led to wide confidence limits for the gross variability. When the number of levels is very small, they will be so wide that it will not be worth spending time estimating them. Thus it is desirable that the number of levels should be as large as possible.

The ideal experiment for this type of analysis will have two replications for each level and a large number of levels, so that both lines of the analysis of variance table are of comparable accuracy. However, if the cost of increasing the number of replications is negligible, there is no harm done if more than two are made. It will be advantageous to make the number of replications reasonably large, if an accurate estimate of the interaction variance is also required from the analysis.

Although the number of levels, K, should be large, experiments are often made in which K is small. For example, in chemical and engineering tests, K has an average of about 10, rarely rising above 20. There are two reasons for this use of a small number of levels of the factor:

(i) The cost of increasing K is usually greater than the cost of increasing the number of replications. An example of this is found in agricultural field trials where the levels of a factor correspond to farms. Here the cost of transporting personnel and equipment to an extra farm will be greater than the cost of taking an extra observation on the farms which have already been chosen for the experiment.

(ii) There may be more trouble involved in increasing K. For example, a manufacturer may send samples of his product to each of K laboratories, n tests being made on the sample in each laboratory. It will be much easier to carry out a few extra tests than to get a larger number of laboratories to co-operate.

However, there is a way in which large values of the degrees of freedom can arise. The experiment can be repeated several times and the results of the separate experiments combined. This, in fact, was done in the example given above. I think that this technique could profitably be applied to other experiments in which high accuracy is required.

8. CONCLUSION

A series expansion suitable for estimating confidence limits for a general linear combination of variances has been derived as far as the term of order $f^{-\frac{3}{2}}$. In numerical work for the case of a positive combination of two variances, it has shown itself to be remarkably good, the best results being obtained at the upper end of the distribution. In view of this fact, tables of the 1, 5, 95 and 99 % points have been calculated and are given in this paper. They are suitable for estimating confidence intervals for the sum of two variances, if the degrees of freedom are large enough. An example of their use is given in the text.

REFERENCES

BARTLETT, M. S. (1953). Approximate confidence intervals. II. More than one unknown parameter. *Biometrika*, **40**, 306–17.

CORNISH, E. A. & FISHER, R. A. (1937). Moments and cumulants in the specification of distributions. *Extrait de la Revue de l'Institute International de Statistique*, **4**, 1–14. Reprinted as paper 30 in *Contributions to Mathematical Statistics*, by R. A. Fisher (1950). John Wiley and Sons, Inc.

TIPPETT, L. H. C. (1952). *Technological Applications in Statistics*. Williams and Norgate Ltd.

WELCH, B. L. (1956). On linear combinations of several variances. *J. Amer. Statist. Ass.* **51**, Part 1.

INTERPOLATIONS AND APPROXIMATIONS RELATED TO THE NORMAL RANGE

By JOHN W. TUKEY*

Princeton University

Many quantities related to the normal range are only well known for a rather open grid of values, and interpolation is often considered difficult (vide David, Hartley & Pearson's (1954) remark at p. 486). Methods of convenient interpolation should thus be of use.

Elfving (1947) established a general asymptotic result for ranges of samples from symmetrical distributions. For the normal distribution this implies that

$$n\int_{\frac{1}{2}w}^{\infty} e^{-\frac{1}{2}x^2}\,dx$$

has a limiting distribution. The usual limiting value of Mill's ratio yields

$$\int_{u}^{\infty} e^{-\frac{1}{2}x^2}\,dx \sim u^{-1}e^{-\frac{1}{2}u^2}.$$

When we combine these and take logarithms, we learn that

$$w^2 + 8\ln w - 8\ln n$$

has a limiting distribution. We might then expect, when the middle term could be neglected, that w^2 would behave like $8\ln n = 18{\cdot}42\log_{10} n$, while the reciprocal of the variance of w would be proportional to

$$\left(\frac{d(w^2+8\ln w)}{dw}\right)^2 = 4w^2\left(1+\frac{4}{w^2}\right)^2 \sim 18{\cdot}42\log_{10} n + 8.$$

Thus the use of $$f_{a,b,c}(n) = a\log_{10} b(n+c)$$

would seem to offer promise as a basis for interpolation.

With suitable choices of b and c, the ratios of suitable powers of:

(1) $w_{5\%}$ and $w_{1\%}$, the upper 5 and 1 % points of the normal range,
(2) d_n and V_n, the average value and variance of the normal range,
(3) the ratio $d_n/\sqrt{V_n}$, and
(4) $(w/s)_{5\%}$ and $(w/s)_{1\%}$, the upper 5 and 1% points of the ratio of a range to s from the same sample,

to $\log_{10} b(n+c)$ are remarkably constant. This is shown in Table 1, where the values of $n > 20$ are those most likely to be key values in view of the existence of standard deviations for the normal range for these values. The ratios have been carried to enough decimal places to show either trends in the ratio, or irregularities which reflect limitations in the available tables.

The constancy of the ratios is so great as to provide not only interpolations but simple and rather accurate approximations. With the growing use of internally programmed automatic

* Prepared in connexion with research sponsored by the U.S. Office of Naval Research.

computers, the already substantial convenience of such simple analytical approximations is being substantially enhanced.

In view of the surprisingly good approximations obtained by choosing b and c properly, it is important to emphasize that useful interpolation is also quite possible and even simpler, since division by the square root of $\log_{10} n$ will make almost all of the quantities treated easily interpolable for n between 20 and 1000.

The 1 % points of the range for $n \leqslant 20$ were taken from Table 22 of the new *Biometrika* tables (based in part on Pearson (1942)), while the corresponding 5 % points were interpolated in Table 23 to give a third decimal, values for $30 \leqslant n \leqslant 100$ were taken from Pearson (1932), while values for $n = 200$, 500 and 1000 were computed from the four moments found by Tippett (1925), using Table 42 of the *Biometrika* tables (based in part on Pearson & Merrington (1951)). The values so found were:

n	100	200	500	1000
$w_{5\%}$	6·09	6·49	7·00	7·37
$w_{1\%}$	6·62	7·02	7·50	7·84

(the values for $n = 100$ being included for comparison with Pearson's values of 6·08 and 6·63). The approximations (corresponding to '*std*' in Table 1)

$$w_{5\%} = \{17{\cdot}0 \log_{10}(1{\cdot}5n)\}^{\frac{1}{2}},$$
$$w_{1\%} = \{17{\cdot}5 \log_{10}(3{\cdot}3n)\}^{\frac{1}{2}},$$

seem to be about as accurate as the tabulated values for $20 \leqslant n \leqslant 1000$ and to be satisfactory for $n \geqslant 6$. For some purposes, it is more convenient to write the approximations in the form (corresponding to '*alt*' in Table 1)

$$(w_n)_{5\%}/(w_2)_{5\%} = 1{\cdot}49\{\log_{10}(1{\cdot}5n)\}^{\frac{1}{2}},$$
$$(w_n)_{1\%}/(w_2)_{1\%} = 1{\cdot}51\{\log_{10}(3{\cdot}3n)\}^{\frac{1}{2}}.$$

The average values of the normal range, d_n, were taken from Table 27 of the *Biometrika* tables (based on Tippett (1925)). The approximation ('*std*' in Table 1)

$$d_n = \{17{\cdot}06 \log_{10} 0{\cdot}291(n + 2{\cdot}6)\}^{\frac{1}{2}}$$

is good to about 1 part in 1500 for $20 \leqslant n \leqslant 1000$, while interpolation in Table 1 for a more precise ratio than 17·06 will easily give an accuracy of 1 in 20,000. The alternate approximation, which in this case agrees with the approximation just given to 4 significant figures, is

$$d_n/d_2 = 3{\cdot}66\{\log_{10} 0{\cdot}291(n + 2{\cdot}6)\}^{\frac{1}{2}}.$$

The variances of the normal range, V_n, were taken from Table 20 of the *Biometrika* tables (based on Hartley & Pearson (1951)) for $n \leqslant 20$, from Tippett (1925), as repeated on p. 45 of the *Biometrika* tables, for $n = 60$, 100, 200, 500 and 1000, and from Pearson (1932) for $n = 30$, 45 and 75. The approximation

$$1/V_n = 1{\cdot}338 \log_{10} 1{\cdot}05(n + 4{\cdot}4)$$

is apparently about as good as the tabulated values for $20 \leqslant n \leqslant 1000$ and is quite useful for $n \geqslant 6$. The alternate form differs by about 2 parts in 1300 and is

$$V_n/V_2 = 1{\cdot}03/\log_{10} 1{\cdot}05(n+4{\cdot}4).$$

Table 1. *Ratios of various quantities to* $\log_{10}(b(n+c))$ *for selected values of n*

Quantity	$[w_{5\%}]^2$	$[w_{1\%}]^2$	d_n^2	$d_n/\sqrt{V_n}$	$d_n/\sqrt{V_n}$	$1/V_n$	$[(w/s)_{5\%}]^2$	$[(w/s)_{1\%}]^2$
b	1·5	3·3	0·291	0·52	0·494	1·05	0·78	1·1
c	0·0	0·0	2·6	2·5	$1{\cdot}535\sqrt[4]{n}$	4·4	−4·5	−7·4
$n=2$	16·10	16·17	10·0	3·58	6·84	1·66	Imag.	Imag.
3	16·82	17·05	13·5	4·18	4·900	1·423	Imag.	Imag.
4	16·9	17·28	15·0	4·42	4·8340	1·366	Imag.	Imag.
5	17·000*	17·38	15·69	4·56	4·8347	1·3440	Imag.	Imag.
6	17·035*	17·47	16·12	4·630	4·8349	1·3343	133·	Imag.
8	17·037*	17·52	16·57	4·711	4·8351	1·3348	39·	86·
10	17·026*	17·53	16·78	4·7497	4·8351	1·3340	21·	32·
15	17·041*	17·53	16·99	4·7871	4·8353	1·3359	19·04	21·3
20	17·006*	17·54	17·054	4·7984	4·8352	1·3370	18·62	20·10
30	17·06	17·50	17·082	4·8034	4·8345	1·3378	(18·41)	(19·73)
45	16·96	17·47	17·079	4·8029	4·8341	1·3382	*n/a*	*n a*
60	16·98	17·50	17·070	4·8015	4·8343	1·3382	18·49	19·95
75	17·03	17·49	17·062	4·8001	4·8336	1·3377	*n/a*	*n/a*
100	16·99†	17·45‡	17·052	4·8008	4·8355	1·3393	18·59	20·14
200	17·00	17·43	17·036	4·7979	4·8345	1·3388	18·64	20·17
500	17·04	17·48	17·044	4·7951	4·8317	1·3370	18·62	20·14
1000	17·10	17·49	17·056	4·8008	4·8358	1·3396	18·59	20·03
std	17·0	17·5	17·06	4·8	4·835	1·3380	18·6	20·1
alt	17·5	17·52	17·06	—	—	1·336	18·49	20·25

* Based on interpolation in the cumulative distribution to a 3rd decimal.
† 17·04 by method used for $n > 100$.
‡ 17·40 by method used for $n > 100$.
Imag. = imaginary, n/a = no percentage point tabulated.
Values in ( ) based on interpolations by Pearson (1932).

Two sorts of approximations to the reciprocal of the coefficient of variation of the range are given. The first yields

$$d_n/\sqrt{V_n} = 4{\cdot}8\log_{10} 0{\cdot}52(n+2{\cdot}5),$$

and is accurate to the accuracy of the known values or to 1 part in 1000 for $n \geqslant 20$. The second, which illustrates an altered form for c, yields

$$d_n/\sqrt{V_n} = 4{\cdot}835\log_{10} 0{\cdot}494(n+1{\cdot}535\sqrt[4]{n}),$$

which is accurate to the accuracy of the known values or to 1 part in 4000 for $n \geqslant 4$. (No alternatives are provided.)

The percentage points of the ratio of a range to the standard-deviation estimate, s, from the same sample were taken from the recent paper by David *et al.* (1954). The approximations

$$(w/s)_{5\%} = \{18{\cdot}6 \log_{10} 0{\cdot}78(n-4{\cdot}5)\}^{\frac{1}{2}},$$
$$(w/s)_{1\%} = \{20{\cdot}1 \log_{10} 1{\cdot}1(n-7{\cdot}4)\}^{\frac{1}{2}},$$

are not quite as good as the key values, the ratios in Table 1 showing a tendency to run low near $n = 50$, and, probably, high near $n = 300$. The amount of this effect is near the limit of accuracy, and should not amount to more than about 1 part in 400 for $20 \leqslant n \leqslant 1000$. About as much more error is introduced by two-significant-figure constants in the alternate forms

$$(w/s)_{5\%} = 4{\cdot}3\{\log_{10} 0{\cdot}78(n-4{\cdot}5)\}^{\frac{1}{2}},$$
$$(w/s)_{1\%} = 4{\cdot}5\{\log_{10} 1{\cdot}1(n-7{\cdot}4)\}^{\frac{1}{2}}.$$

METHODS

The approximations given are not best in any specific sense, but have been chosen to fit reasonably well. Except for $w_{5\%}$ and $w_{1\%}$, where $c = 0$ works quite well, the values of b and c were usually determined to make the ratios for $n = 20, 100$ and 1000 nearly the same, which can be easily done by trial and error solution of

$$\log_{10}(100+c) - \frac{\log_{10}(1000+c) - \log_{10}(100+c)}{(u_{1000}/u_{100}) - 1}$$
$$= \log_{10}(20+c) - \frac{\log_{10}(1000+c) - \log_{10}(20+c)}{(u_{1000}/u_{20}) - 1},$$

the common value of these differences being taken as $-\log_{10} b$.

ILLUSTRATION OF APPROXIMATION

To illustrate the use of the technique in more refined approximation, Fig. 1 shows values of

$$\text{antilog}_{10}(d_n/4{\cdot}835\sqrt{V_n}) - 0{\cdot}496n - 0{\cdot}6\log_{10} n$$

plotted against $\log_{10} n$. For larger values of n the dashed vertical lines indicate the uncertainties associated with the fact that V_n is given to only 3 decimals for $n = 60, 100, 200, 500$ and 1000, and to only 4 decimals for 30, 45 and 75. Interpolation to nearly 4-decimal accuracy in V_n is clearly possible from $n = 20$ nearly to $n = 100$. Better values for n's between 100 and 1000 would clearly be obtained from one or two 4 or 5 decimal values of V_n, rather than by many more values of lower accuracy. A single highly accurate quadrature for an n of, say, 1001 would probably go far to settle this situation.

The quantity plotted can be regarded as

$$0{\cdot}496c' - 0{\cdot}6\log_{10} n,$$

where

$$d_n/\sqrt{V_n} = 4{\cdot}835\log_{10} 0{\cdot}496(n+c').$$

Here 4·835 was chosen from the second approximation to $d_n/\sqrt{V_n}$, while 0·496 was modified from 0·494 to give a relatively straight plot for n between 10 and 100. The use of $0{\cdot}6\log_{10} n$ is not essential, and serves merely to simplify the plot somewhat.

A further use of the same technique is in supplementing the available values of the lower percentage points of w/s. The lower 2·5 % point serves as a good example. Fitting to $n = 10$. 100 and 1000 yields

$$3{\cdot}907 + \{\log_{10}[(n + 11{\cdot}42)/7{\cdot}79]\}^{\frac{1}{2}}$$

as a first approximation. Keeping the 3·907, and adjusting the 7·79 to 8·14 to produce reasonable linearity against $\log n$, leads to the use of

$$z = 8{\cdot}14\,\text{antilog}_{10}\,(y/3{\cdot}907)^2 - n - 4{\cdot}6\log_{10} n,$$

as a quantity of reasonable constancy, where y is the lower 2·5 % point of w/s.

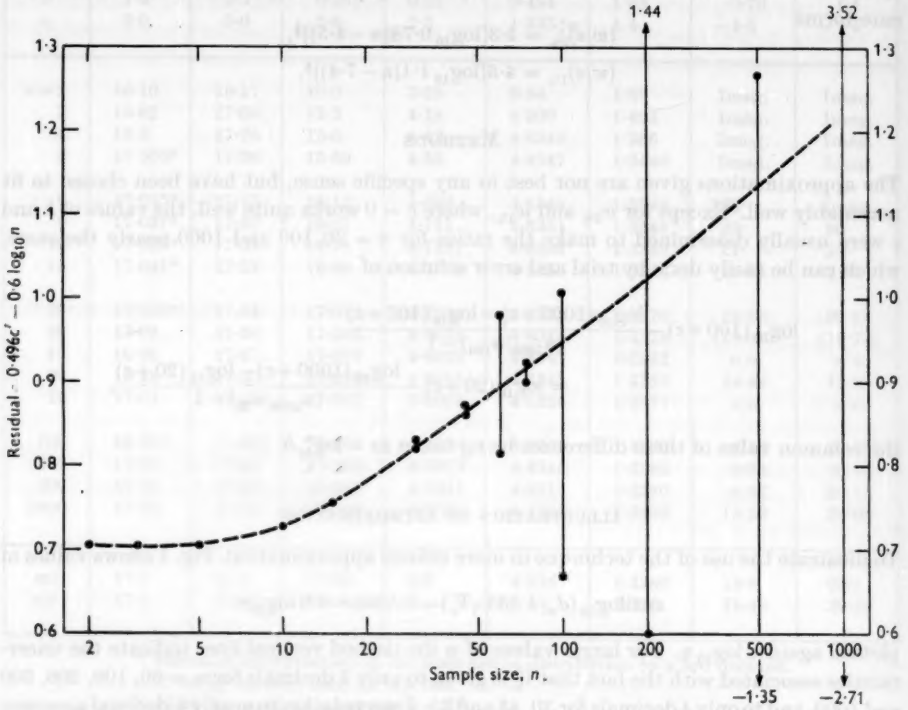


Fig. 1. Residuals for $d_n/\sqrt{V_n}$, where c' is defined by $d_n/V_n = 4{\cdot}835\ \log_{10} 0{\cdot}496\ (n + c')$.

The numerical values found are as follows:

n	2	3	10	15	20	60	100	200	500	1000
z	7·63	7·69	7·79	8·00	8·70	7·97	7·28	7·87	7·89	43·42

The values for $n = 2$ and 3 are taken from Thomson (1955). There appears to be some ground for suspicion that the values of y for 20 and 1000 are 0·02 or 0·03 high. Interpolation for $n = 4$ to 9 should be simple and effective.

While the original approximation (to $w_{5\%}$) was obtained entirely empirically, the writer owes the suggestion of its asymptotic explanation to his former colleague, David L. Wallace.

REFERENCES

DAVID, H. A., HARTLEY, H. O. & PEARSON, E. S. (1954). *Biometrika*, **41**, 482–93.
ELFVING, G. (1947). *Biometrika*, **34**, 111–19.
HARTLEY, H. O. & PEARSON, E. S. (1951). *Biometrika*, **38**, 463.
PEARSON, E. S. (1932). *Biometrika*, **24**, 404–17.
PEARSON, E. S. (1942). *Biometrika*, **32**, 301–8.
PEARSON, E. S. & MERRINGTON, M. (1951). *Biometrika*, **38**, 4–10.
THOMSON, G. W. (1955). *Biometrika*, **42**, 268–9.
TIPPETT, L. H. C. (1925). *Biometrika*, **17**, 364–87.

THE GAMBLER'S RUIN PROBLEM WITH CORRELATION

By C. MOHAN

London School of Economics and Columbia University, New York

1. Introduction

In the classical problem of the gambler's ruin there is a constant probability of the gambler winning any game and the games are independent. Feller, for example, has treated this problem in his book (1950). We consider here a more general problem where there is a correlation between the results of two successive games. We consider two players X, Y. Conditional on winning the previous game, the probabilities that X wins or loses the next game are p_1, q_1, and conditional on losing the previous game these probabilities are q_2, p_2. Thus X's ability to win or lose is governed by the transition probability matrix (t.p.m.):

$$\begin{array}{c} \\ \text{win} \\ \text{loss} \end{array} \begin{array}{c} \begin{array}{cc} \text{win} & \text{loss} \end{array} \\ \begin{pmatrix} p_1 & q_1 \\ q_2 & p_2 \end{pmatrix} \end{array}, \tag{1·1}$$

where

$$p_1+q_1 = 1 = p_2+q_2. \tag{1·2}$$

If we suppose that the stake in each game is one coin, two successive games can have for X one of the following pairs of results, the transition probabilities being given before the colon:

$$p_1\colon\ 1\to 1, \quad q_1\colon\ 1\to -1, \quad p_2\colon\ -1\to -1, \quad q_2\colon\ -1\to 1.$$

On the assumption that a score of $+1$ has the same probability as a score of -1 in the first game, the coefficient of correlation is

$$R = \frac{p_1-q_2}{\sqrt{(p_1+q_2)(p_2+q_1)}}. \tag{1·3}$$

When $p_1 = p_2 = p$ (so that also $q_1 = q_2 = 1-p = q$), the correlation becomes $R = p-q$. This corresponds to a real situation, as, for example, a game between two players, of whom the one holding the 'bank' has an advantage, the bank passing to the winner of each game. We suppose that the probability of winning with the bank is p, and $p>q$, so that $R>0$. On the other hand, if the rule were that the bank passes to the loser of each game, we should have a negative correlation R.

Suppose that X starts with a capital r and Y with a capital $\alpha-r$, and the games are played with a certain (positive or negative) correlation R, until one or other of the players is ruined. How is the probability of ruin affected by R? In particular, if X had the choice of determining whether the games were to be played with a positive or negative correlation, on what basis should he decide so as to maximize his chances of winning?

This problem is analogous to a correlated random walk in the presence of absorbing barriers at $n = 0$ and at $n = \alpha$, the particle starting from a position $r(0<r<\alpha)$ and moving a unit step to the right or to the left according to the t.p.m. (1·1).

The random walk is said to be symmetric if $p_2 = p_1 = p$. Then at the end of each step there is a probability p of a step being in the same direction as the preceding one and a

probability $q\ (=1-p)$ of the direction being reversed. We shall consider first the problem for the general unsymmetric case corresponding to the t.p.m. (1·1).

In case the correlation is zero (i.e. $p_1 = q_2 = \overline{p}$, so that $q_1 = p_2 = 1-\overline{p}$), the steps are to the right with a constant probability $\overline{p}$ and to the left with a constant probability $1-\overline{p}$; thus there is a drift towards the right which is measured by $2\overline{p}-1$. When $\overline{p} = \frac{1}{2}$ the random walk is both symmetric and uncorrelated.

In the case of unsymmetrical correlated walk, the drift may still be measured by the quantity

$$\delta = p_1 - p_2.$$

Introducing the quantity

$$c = p_1 + p_2 - 1, \tag{1·4}$$

the coefficient of correlation R given by (1·3) is now given in terms of c and δ by the relation

$$R^2(1-\delta^2) = c^2. \tag{1·5}$$

It will be noticed that in the symmetrical case ($\delta = 0$) $c = R$.

2. Probability of ultimate ruin

Denote by a_r and b_r the probabilities of ultimate ruin of X (whose initial capital is r) conditional on the first game being a win or a loss. It is easy to see that a_r and b_r satisfy the recurrence relations

$$\left.\begin{aligned} a_r &= p_1 a_{r+1} + q_1 b_{r+1} \\ \text{and} \quad b_{r+1} &= p_2 b_r + q_2 a_r \end{aligned}\right\} \quad (0 < r < \alpha - 1). \tag{2·1}$$

Denoting by E the operation of increasing the suffix r by unity, these may be written as

$$(1-p_1E)\,a_r = q_1 E b_r,$$
$$(E-p_2)\,b_r = q_2 a_r$$

which show that a_r and b_r satisfy the difference equation

$$(E-p_2)(1-p_1E)\,\chi_r = q_1 q_2 E \chi_r,$$

i.e.

$$\left\{E^2 - \frac{1+c}{p_1}E + \frac{p_2}{p_1}\right\}\chi_r = 0.$$

Hence

$$b_r = G + H\left(\frac{p_2}{p_1}\right)^r,$$

$$a_r = \frac{1}{q_2}\left[G + H\left(\frac{p_2}{p_1}\right)^{r+1} - p_2\left\{G + H\left(\frac{p_2}{p_1}\right)^r\right\}\right].$$

The boundary values are $b_1 = 1$ and $a_{\alpha-1} = 0$; whence

$$G + H\left(\frac{p_2}{p_1}\right) = 1, \quad q_2 G + q_1\left(\frac{p_2}{p_1}\right)^{\alpha} H = 0,$$

so that

$$p_2^{\alpha} q_1 / G = -q_2 p_1^{\alpha} / H = p_2(p_2^{\alpha-1} q_1 - q_2 p_1^{\alpha-1}).$$

Substituting these values of G and H we are led to the probabilities of ultimate ruin

$$\left.\begin{aligned} b_r &= \frac{q_1\lambda^{\alpha} - q_2\lambda^r}{q_1\lambda^{\alpha} - q_2\lambda} \cdot \\ \text{and} \quad a_r &= \frac{q_1\lambda^{\alpha} - q_1\lambda^{r+1}}{q_1\lambda^{\alpha} - q_2\lambda}, \end{aligned}\right\} \tag{2·2}$$

where $\lambda = p_2/p_1$.

Now if c_1 denotes the probability of initial win and c_2 $(= 1-c_1)$ of initial loss, then the general expression for P_r, the probability of ultimate ruin unconditional on the result of the first game, is

$$P_r = \frac{c_1 q_1(\lambda^{\alpha-1}-\lambda^r)+c_2(q_1\lambda^{\alpha-1}-q_2\lambda^{r-1})}{(q_1\lambda^{\alpha-1}-q_2)}. \tag{2·3}$$

For $\lambda = 1$, (2·3) is indeterminate. In this symmetrical case the equations (2·1) reduce to

$$a_r = pa_{r+1}+qb_{r+1}$$

and

$$b_{r+1} = pb_r+qa_r.$$

Hence, it is found that P_r is now given by

$$P_r = \frac{c_2+(\alpha-1-r)q}{1+q(\alpha-2)}. \tag{2·4}$$

P_r can be expressed in terms of δ and c, since

$$\lambda = (c+1-\delta)/(c+1+\delta) \quad \text{and} \quad q_2/q_1 = (1-c+\delta)/(1-c-\delta).$$

As the formula in its general form is not easy to interpret, we take for illustration the symmetrical case

$$p_2 = p_1 = p \quad \text{with} \quad c_1 = c_2 = \tfrac{1}{2}, \tag{2·5}$$

and study the behaviour of P_r as the correlation R varies from -1 to $+1$. Now

$$R = p-q = 1-2q = c.$$

Let $\rho = 2R/(1-R)$, so that $q = \frac{1}{2}(1-R) = 1/(\rho+2)$. Then from (2·4)

$$P_r = \frac{1}{2}\left(1+\frac{1-2\frac{r}{\alpha}}{1+\frac{\rho}{\alpha}}\right). \tag{2·6}$$

Now ρ varies between -1 and $+\infty$ as R varies between -1 to $+1$, and $\alpha > r > 0$. Hence

$$P_r \text{ lies between } \tfrac{1}{2} \text{ and } 1-(r-\tfrac{1}{2})/(\alpha-1).$$

Thus $P_r \gtrless \frac{1}{2}$ according as $r \lesseqgtr \frac{1}{2}\alpha$. It will be noticed that while $\lim_{R\to 1} P_r = \frac{1}{2}$, $\lim_{R\to -0} P_r$ is different in the three cases $r \lesseqgtr \frac{1}{2}\alpha$. The situation is shown below in the graphs of equation (2·6) for three typical cases. Each curve cuts off from the vertical axis an intercept $(1-r/\alpha)$; as $R\to+1$, $P_r\to\frac{1}{2}$; when $R\to-1$, $P_r\to 1-(r-\frac{1}{2})/(\alpha-1)$. These results are easy to interpret, for when $R\to 1$ the probability of ultimate ruin depends less and less on the initial capital and should more and more approach the probability of losing the very first game; this is shown by the three curves approaching one another as $R\to 1$. The value of P_r for $R=-1$ is actually non-existent, as is borne out by a study of the equations (2·1) which for $R=-1$ degenerate into a single set of equations

$$a_r = b_{r+1} \quad (1 \leqslant r \leqslant \alpha-2),$$

with the boundary values $a_{\alpha-1} = 0$, $b_1 = 1$. These $(\alpha-2)$ equations are not soluble; in fact, these show that $a_1, a_2 \ldots a_{\alpha-2}$ are respectively equal to $b_2, b_3, \ldots, b_{\alpha-1}$ and none of them involves either of the unknown probabilities $a_{\alpha-1}, b_1$. On the other hand, when $R=1$, equations (2·1) become

$$a_r = a_{r+1}, \quad b_r = b_{r+1} \quad (1 \leqslant r \leqslant \alpha-2),$$

with the boundary values $b_1 = 1$, $a_{\alpha-1} = 0$. Hence

$$a_1 = a_2 = \ldots = a_{\alpha-1} = 0 \quad \text{and} \quad b_1 = b_2 = \ldots = b_{\alpha-1} = 1,$$

so that $$P_r = c_1 a_r + c_2 b_r = c_2 \quad (= \tfrac{1}{2} \text{ for the graph}).$$

We conclude that if X's capital is initially less than Y's, it is advantageous for him to choose the positively correlated play, and if initially greater than Y's, then to choose negatively correlated play. In each case the greater the correlation in magnitude the more it is to be preferred.

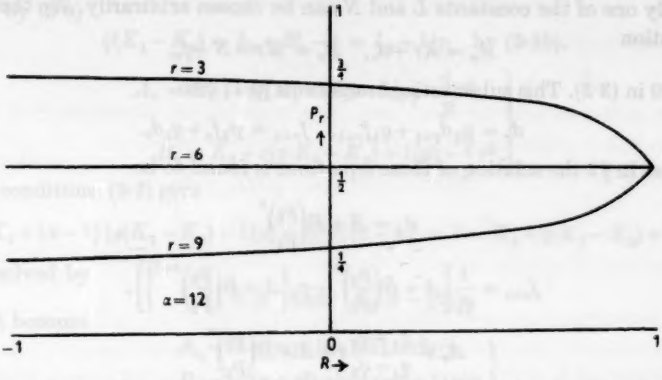


Fig. 1. The graphs show the probability of ruin, P_r, as a function of the correlation R in a game in which the total capital is 12 for three cases of initial capitals of $r = 3$, 6 and 9.

3. Expected duration of play

The expected value of the duration of play can also be determined by the use of difference equations. The method is justified if we assume that these expectations are finite. Let A_r, B_r be the respective durations of play if the initial capital is r and the first game results in a win or loss.

Then if the first game results in a win the capital becomes $r+1$ and then the expected duration is A_{r+1} with probability p_1 and B_{r+1} with probability q_1. Hence

$$\left.\begin{aligned} A_r &= p_1 A_{r+1} + q_1 B_{r+1} + 1 \\ \text{similarly} \quad B_{r+1} &= p_2 B_r + q_2 A_r + 1 \end{aligned}\right\} \quad (0 < r < \alpha - 1). \tag{3·1}$$

The boundary values are

$$B_1 = 1 = A_{\alpha-1}. \tag{3·2}$$

Equations (3·1) become homogeneous on putting

$$A_r = Kr + L + d_r, \quad B_r = Mr + N + f_r \tag{3·3}$$

if
$$Kr + L = p_1(Kr + K + L) + q_1(Mr + M + N),$$
$$Mr + M + N = p_2(Mr + N) + q_2(Kr + L) + 1;$$

i.e. if
$$K = p_1 K + q_1 M, \qquad L = p_1(K + L) + q_1(M + N) + 1,$$
$$M = p_2 M + q_2 K, \quad M + N = p_2 N + q_2 L + 1.$$

Remembering the relation (1·2), viz. $p_1+q_1 = 1 = p_2+q_2$, these equations are solved by

$$K = M, \quad q_1(L-N) = M+1 \quad \text{and} \quad q_2(L-N) = M-1,$$

i.e.
$$K = M = (q_1+q_2)/(q_1-q_2) \quad \text{and} \quad L-N = 2/(q_1-q_2). \tag{3·4}$$

This solution shows that the homogeneity of (3·1) achieved by the substitution (3·3) is not disturbed if the same constant is added to both the relations (3·3). This conclusion is corroborated by a study of the structure of equations (3·1); for if A_r and B_r are increased by k (a constant with regard to r) then by virtue of the relation (1·2) k cancels out from them. Consequently one of the constants L and N can be chosen arbitrarily. We therefore make the substitution

$$A_r = Kr+d_r, \quad B_r = Mr+N+f_r,$$

i.e. set $L = 0$ in (3·3). This substitution transforms (3·1) into

$$d_r = p_1 d_{r+1}+q_1 f_{r+1}, \quad f_{r+1} = p_2 f_r+q_2 d_r.$$

Proceeding as in § 2 the solution of these equations is found to be

$$d_r = A+B\left(\frac{p_2}{p_1}\right)^r,$$

$$f_{r+1} = \frac{1}{q_1}\left[A+B\left(\frac{p_2}{p_1}\right)^r - p_1\left\{A+B\left(\frac{p_2}{p_1}\right)^{r+1}\right\}\right],$$

so that
$$A_r = \frac{q_1+q_2}{q_1-q_2}r+A+B\left(\frac{p_2}{p_1}\right)^r,$$

and
$$B_{r+1} = \frac{q_1+q_2}{q_1-q_2}(r+1)-\frac{2}{q_1-q_2}+A+B\left(\frac{p_2}{p_1}\right)^r\frac{q_2}{q_1}.$$

The boundary values (3·2) give

$$\frac{q_1+q_2}{q_1-q_2}(\alpha-1)+A+B\left(\frac{p^2}{p_1}\right)^{\alpha-1} = 1 = \frac{-2+q_1+q_2}{q_1-q_2}+A+B\frac{q_2}{q_1},$$

whence
$$B = \frac{(\alpha-2)(q_1+q_2)+2}{(q_2/q_1-\lambda^{\alpha-1})(q_1-q_2)}, \quad A = \frac{-2q_2+2}{(q_1-q_2)}-\left(\frac{q_2}{q_1}\right)B.$$

Thus
$$A_r = \frac{(q_1+q_2)r+2p_2}{q_1-q_2}+\frac{(\alpha-2)(q_1+q_2)+2}{(q_1-q_2)}\frac{\lambda^r-q_2/q_1}{q_2/q_1-\lambda^{\alpha-1}},$$

and
$$B_r = \frac{(q_1+q_2)r-2q_2}{q_1-q_2}+\frac{(\alpha-2)(q_1+q_2)+2}{(q_1-q_2)}\frac{q_2}{q_1}\frac{\lambda^r-1}{q_2/q_1-\lambda^{\alpha-1}}. \tag{3·5}$$

We now consider the symmetrical case $q_1 = q_2$. The relations (3·4) are indeterminate in this case. We therefore try the quadratic expressions

$$A_r = K_1+L_1r+M_1r^2, \quad B_r = K_2+L_2r+M_2r^2.$$

Substituting these expressions in

$$A_r = pA_{r+1}+qB_{r+1}+1, \quad B_{r+1} = pB_r+qA_r+1, \tag{3·6}$$

and identifying the various powers of r,

$$K_1 = p(K_1+L_1+M_1)+q(K_2+L_2+M_2)+1,$$

$$L_1 = p(L_1+2M_1)+q(L_2+2M_2), \quad M_1 = pM_1+qM_2,$$

$$K_2+L_2+M_2 = qK_1+pK_2+1, \quad L_2+2M_2 = qL_1+pL_2, \quad M_2 = qM_1+pM_2.$$

The third and sixth give $M_1 = M_2$ and the other four are

$$q(K_2-K_1)+pL_1+qL_2+M_1+1=0, \tag{3·7}$$

$$q(L_2-L_1)+2M_1=0, \tag{3·8}$$

$$q(K_1-K_2)-L_2-M_1+1=0, \tag{3·9}$$

and one identical with (3·8).

Adding (3·7) and (3·9)

$$p(L_1-L_2)+2=0, \tag{3·10}$$

so that by (3·8) $M_1 = \frac{1}{2}q(L_1-L_2) = -q/p,$

and finally by (3·9)

$$q(K_1-K_2) = L_2+M_1-1 = L_2+1/p \quad \text{by (3·10).}$$

Thus

$$\left.\begin{aligned} A_r &= K_1+r\{q(K_1-K_2)-1/p\}-\frac{q}{p}r^2, \\ B_r &= K_2+r\{q(K_1-K_2)+1/p\}-\frac{q}{p}r^2. \end{aligned}\right\} \tag{3·11}$$

The initial conditions (3·2) give

$$K_1+(\alpha-1)\{q(K_1-K_2)-1/p\}-(q/p)(\alpha-1)^2 = 1 = K_2+q(K_1-K_2)+1,$$

which are solved by

$$K_1 = \alpha \quad \text{and} \quad K_2 = -q\alpha/p.$$

Thus (3·11) become

$$\left.\begin{aligned} A_r &= \alpha+(r/p)(\alpha q-1)-r^2q/p, \\ \text{and} \quad B_r &= -(\alpha+r^2)q/p+(\alpha q+1)r/p. \end{aligned}\right\} \tag{3·12}$$

Hence the expected duration of play unconditional upon the result of the first game is

$$D_r = c_1\left\{(\alpha-r)+\frac{q}{p}r(\alpha-1-r)\right\}+c_2\left\{r+\frac{q}{p}(r-1)(\alpha-r)\right\}, \tag{3·13}$$

where c_1, c_2 are, as already defined, the initial probabilities of win and loss. Assuming the first win and loss to be equally probable (3·13) becomes

$$D_r = r(\alpha-r)-\frac{R}{1+R}\{-\alpha+2r(\alpha-r)\}. \tag{3·14}$$

When $R=0$, $D_r = r(\alpha-r)$, in agreement with Feller's formula (1950, p. 287). Further, since $r(\alpha-r) \geqslant \frac{1}{4}\alpha^2 \geqslant \frac{1}{2}\alpha$ (as $\alpha \geqslant 2$), it follows that except in the trivial case $\frac{1}{2}\alpha = 1 = r$, for positively correlated play the duration has smaller expectation than for the uncorrelated case and for negatively correlated play it has larger expectation. In fact as $R \to 1$, $D_r \to \frac{1}{2}\alpha$ and as $R \to -1$, $D_r \to \infty$.

4. Generating functions for the probabilities of the gambler's ruin at the nth game

Denote by $v_{r,n}$ the conditional probability that X (whose initial capital is r) is ruined at the nth game, given that he wins the first game and by $w_{r,n}$ the corresponding probability when he loses the first game.

Evidently $v_{r,n} = 0$ if $r > n-2$ or if $n-r$ is odd,

and $w_{r,n} = 0$ if $r > n$ or if $n-r$ is odd. (4·1)

Define the generating functions (g.f.) of $v_{r,n}$, $w_{r,n}$

$$V_r(s) = \sum_{n=1}^{\infty} v_{r,n} s^n, \quad W_r(s) = \sum_{n=1}^{\infty} w_{r,n} s^n. \tag{4·2}$$

Elementary reasoning shows that the following recurrence relations hold

$$\left.\begin{aligned} v_{r,n+1} &= p_1 v_{r+1,n} + q_1 w_{r+1,n} \\ w_{r+1,n+1} &= p_2 w_{r,n} + q_2 v_{r,n} \end{aligned}\right\} \quad (1 < r < \alpha - 1,\ n \geqslant 1). \tag{4·3}$$

Evidently
$$w_{1,1} = 1, \quad v_{\alpha-1,n} = 0, \quad n \geqslant 1. \tag{4·4}$$

Define the boundary value
$$w_{1,n} = 0 \quad (n > 1) \tag{4·5}$$

in order that (4·3) may hold for $r = 1$ also. Now multiply the equations (4·3) by s^{n+1} and add from $n = 1$ to ∞, then since $v_{r\,1} = 0 = w_{r+1,1}$, $r \geqslant 1$, we have

$$\left.\begin{aligned} V_r(s) &= p_1 s V_{r+1}(s) + q_1 s W_{r+1}(s), \\ \text{and} \qquad W_{r+1}(s) &= p_2 s W_r(s) + q_2 s V_r(s). \end{aligned}\right\} \tag{4·6}$$

Using the operator E already defined these may be written as

$$(1 - p_1 s E) V_r(s) = q_1 s E W_r(s),$$
$$(E - p_2 s) W_r(s) = q_2 s V_r(s),$$

which show that $V_r(s)$ and $W_r(s)$ satisfy the difference equation

$$(E - p_2 s)(1 - p_1 s E) \chi_r = q_1 q_2 s^2 E \chi_r,$$

i.e.
$$\left\{E^2 - \frac{1 + cs^2}{p_1 s} E + \frac{p_2}{p_1}\right\} \chi_r = 0.$$

Hence
$$W_r(s) = A\eta_1^r + B\eta_2^r,$$
so that by (4·6)
$$V_r(s) = (q_1 s)^{-1}\{A\eta_1^{r+1} + B\eta_2^{r+1} - p_2 s(A\eta_1^r + B\eta_2^r)\}, \tag{4·7}$$

where η_1, η_2 are the roots of the quadratic

$$p_1 \eta^2 - \frac{(1 + cs^2)}{s} \eta + p_2 = 0.$$

Since (4·6) are homogeneous in $V_r(s)$, $W_r(s)$ and hold for $r = 1, 2, \ldots, \alpha - 2$, they can determine $V_r(s)$ and $w_r(s)$ for $r = 1, 2, \ldots, \alpha - 1$ if two of these $2(\alpha - 1)$ functions are known. These two known functions are $V_{\alpha-1}(s)$ and $W_1(s)$, since on using (4·2) the relations (4·4) and (4·5) lead to the boundary conditions

$$V_{\alpha-1}(s) = \sum_{n=1}^{\infty} v_{\alpha-1,n} s^n = 0, \quad W_1(s) = \sum_{n=1}^{\infty} w_{1,n} s^n = s. \tag{4·8}$$

For these values (4·7) give

$$A\eta_1 + B\eta_2 = s, \quad A\eta_1^{\alpha-1}(\eta_1 - p_2 s) + B\eta_2^{\alpha-1}(\eta_2 - p_2 s) = 0,$$

whence, writing

$$D(s) = D = \eta_1 \eta_2^{\alpha-1}(\eta_2 - p_2 s) - \eta_1^{\alpha-1}\eta_2(\eta_1 - p_2 s),$$

we have
$$DA = s\eta_2^{\alpha-1}(\eta_2 - p_2 s) \quad \text{and} \quad DB = -s\eta_1^{\alpha-1}(\eta_1 - p_2 s).$$

Substituting these values in (4·7) and noting that

$$(\eta_1 - p_2 s)(\eta_2 - p_2 s) = \frac{p_2}{p_1} - \frac{p_2}{p_1}\{1 + s^2(p_1 + p_2 - 1)\} + p_2^2 s^2 = \frac{p_2 q_1 q_2}{p_1} s^2,$$

we have finally

$$\left.\begin{aligned} V_r(s) &= q_1 s^2 D^{-1}(\eta_2^x \eta_1^{r+1} - \eta_1^x \eta_2^{r+1}), \\ \text{and} \quad W_r(s) &= sD^{-1}\{\eta_2^{x-1}\eta_1^r(\eta_2 - p_2 s) - \eta_1^{x-1}\eta_2^r(\eta_1 - p_2 s)\}. \end{aligned}\right\} \tag{4·9}$$

It may be noted that the probabilities a_r, b_r of ultimate ruin calculated in the second section may be determined from these functions $V_r(s)$ and $W_r(s)$. In fact,

$$a_r = \sum_{n=r+2}^{\infty} v_{r,n} = \sum_{n=1}^{\infty} v_{r,n} = V_r(1),$$

and

$$b_r = \sum_{n=1}^{\infty} w_{r,n} = W_r(1).$$

Similarly, we have

$$A_r = V_r'(1) \quad \text{and} \quad B_r = W_r'(1).$$

5. Observations on the case $\alpha = \infty$

(i) *Probability of ultimate ruin*

Evidently when $p_2 > p_1$, i.e. when the drift δ is negative, as $\alpha \to \infty$, a_r and b_r as given in (2·2) tend to unity. Even when $p_2 = p_1$, i.e. the probabilities of a win or loss being continued are equal, the formula (2·2) show that with probability unity, X will be ruined; but when $p_2 < p_1$, then the probability of ultimate ruin, according as the first game is won or lost is

$$\lambda^r \frac{q_1}{q_2} \quad \text{or} \quad \lambda^{r-1}.$$

(ii) *Duration of play*

(*a*) When $p_2 > p_1$, i.e. when the drift δ is negative, as $\alpha \to \infty$, the last terms in (3·5) tend to zero and

$$A_r \to \frac{(q_1 + q_2)r + 2p_2}{q_1 - q_2} \quad \text{and} \quad B_r \to \frac{(q_1 + q_2)r - 2q_2}{q_1 - q_2}.$$

(*b*) When $p_2 < p_1$, i.e. when there is a positive drift δ, as $\alpha \to \infty$, (3·5) show that A_r and $B_r \to \infty$.

(*c*) When $p_2 = p_1$, as $\alpha \to \infty$, (3·12) shows that A_r and $B_r \to \infty$ and hence from (3·13) $D_r \to \infty$, unless $c_1 = 0 = q$. This exceptional case occurs when with probability unity the first game is lost, as also is each succeeding game, so that X is ruined in r games.

In conclusion, I wish to express my gratitude to Dr F. G. Foster, who suggested this problem to me and also supervised the investigation.

REFERENCE

Feller, W. (1950). *An Introduction to Probability Theory and its Applications.* New York: Wiley and Sons.

TABLES FOR SIGNIFICANCE TESTS OF 2×2 CONTINGENCY TABLES

By P. ARMSEN

National Institute for Personnel Research, South African Council for Scientific and Industrial Research

1. Introduction

In a 2×2 tables of the type

	Y	Not Y	
X	a	c	A
Not X	b	d	B
	R	S	N

it is assumed that a, b, c and d are the absolute frequencies resulting from a double dichotomy of N individuals according to the properties X and Y.

The null hypothesis to be tested is:

There is no association between X and Y. If this is true, one would expect in the long run of experiments with the same marginal totals A, B, R, S, to have $a : \bar{b} = A : B$, or which is the same, $N.\bar{a} = A.R$.

The whole 2×2 table can—for fixed marginal totals—be determined by one of the entries. It is arbitrary which one would choose. The probability of the set of frequencies observed and of possible other sets of frequencies which might have been observed can be computed by the exact hypergeometric formula given by Fisher (1941, p. 95).* From this it follows that the possible sets of frequencies can be looked at as events with a completely known one-variate discrete probability distribution. It is bell-shaped or—in some cases—J-shaped, e.g.

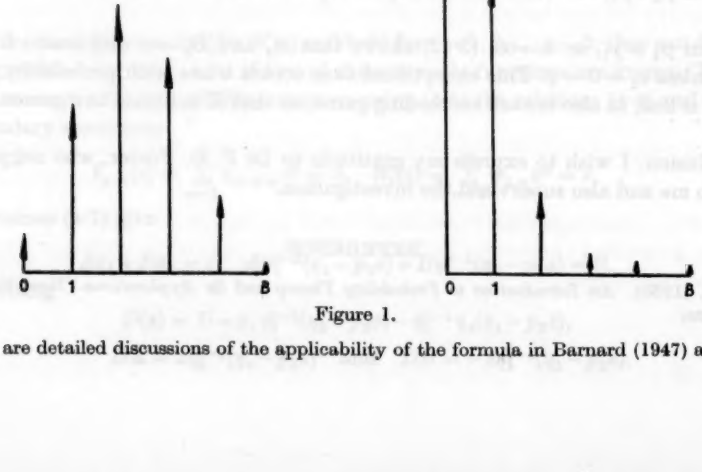


Figure 1.

* There are detailed discussions of the applicability of the formula in Barnard (1947) and Cochran (1952).

The range of the distribution of b or d is from 0 to B, if B is the smallest marginal total.

The present tables are based on direct evaluation of the hypergeometric formula.

It is possible to look at the given null hypothesis in two ways:

(1) The object is to determine experimentally whether or not there is reasonable evidence for a positive association between X and Y (or, which is the same, for a negative association between X and Not Y). In this case it is natural and generally agreed to take the observed table as significant (for departure from the null hypothesis) at, say, the 100α % level, if the sum of the probabilities of the observed and the possible more extreme tables (extreme in the same direction) is smaller than or equal to α, where, for example, $\alpha = 0{\cdot}05$ or $0{\cdot}01$. This is called 'the case of one-tailed significance'.

(2) There is no *a priori* knowledge about the relationship between X and Y. Here a decision must be made before starting the experiment as to which of the possible events (sets of frequencies of 2×2 tables) are to be taken as evidence for an existing association between X and Y (positive or negative whichever the case may be).

In this case we are looking out for alternatives to the null hypothesis which may throw the value of a frequency in the 2×2 table, say d, into either of the tails of the probability distribution. The usual procedure is to cut from each tail a number of terms for which the probability on the null hypothesis sums to $\leqslant \frac{1}{2}\alpha$. If the observed value falls in either of these tails we say that the result is significant at the 100α % level, using a 'two-tailed' test.

In analogous problems where the probability distribution is continuous it is possible to find significance points, whether one-tailed or two-tailed, defining rejection regions associated precisely with any desired value of α. For discontinuous variation which arises when dealing with the hypergeometric type of distribution of our present problem or with the binomial or Poisson distribution, this is not the case.

The position is illustrated in Example 1 given in the following section. If we take $\alpha = 0{\cdot}05$, and look for significance in the sense of a positive relation between X and Y, we shall only establish this if $d = i$ is 6 or 7. But the probability of this result is only $p(6) + p(7) = 0{\cdot}71$ %, far below the 5 % aimed at. If we use a two-tailed test based on cutting off regions from the two ends of the distribution for each of which the sum of the probabilities is $\leqslant 2{\cdot}5$ %, then we see that this region will again only include $i = 6$ and 7.

Unless we adopt a procedure of random sampling for a decision, discussed, for example, by Tocher (1950) and Cochran (1952), there is no means of obtaining exactly 5 % for the rejection region. It will, however, be noted that while for the single-tailed test the region is clearly defined, for the two-tailed test alternative regions are possible if we drop the usual convention of making the probability associated with each tail region sum to $\frac{1}{2}\alpha$ or less. Thus coming back to Example 1, it might be asked whether there is serious objection to including $i = 0$, 6 and 7 in the two-tailed 5 % rejection region, giving a total probability of $2{\cdot}58 + 0{\cdot}71 = 3{\cdot}29$ % (still under 5 %) on the null hypothesis?

The object of this paper is to consider alternative definitions of the rejection regions for the two-tailed test, and also to provide tables of 5 and 1 % significance levels for both one- and two-tailed tests, and for 2×2 tables containing up to $N = 50$ observations. These tables, by adopting a new method of entry, make it possible to go beyond the range of existing tables using relatively little space.

2. Choice of the rejection region for the two-tailed test of significance

The definition of a two-tailed 100α % significance region referred to in the introductory section, which has been used largely, e.g. in Finney's (1948) tables, may be expressed as follows:

Definition D_1. Cut as much as possible but not more than $\frac{1}{2}\alpha \times 100$ % from each tail of the distribution under the null hypothesis.

While in any particular case, when the hypergeometric probabilities have been calculated, it may seem clear on intuitive grounds how to enlarge this rejection region by including one or more further values of d and yet keep the total probability below 100α %, when preparing tables for general use it is necessary to have some clearly defined, consistent rule of procedure.

A possible procedure which the author prefers is given by the following definition which is that used in the mathematical description of the tables (pp. 502–5 below):

Definition D_2. Arrange the possible events (defined say in terms of d) in ascending order of the size of their probabilities under the null hypothesis; include in the 100α %, two-tailed rejection region those events for which the cumulative sum of these ordered probabilities is smaller than or equal to α. (The kind of association—positive or negative—which is likely to exist will be indicated by the observed event.)

An equivalent visual description, which can be illustrated on the diagrams given above, is as follows:

Lift a parallel to the horizontal axis until the sum of the probability ordinates falling below this line has reached as closely as possible, but not exceeded α. Each event whose associated probability ordinate is now completely below the parallel is two-tailed significant at the 100α % level. The others are not.

The advantages of this definition D_2 are that:

(i) It is straightforward to apply.

(ii) It will often include more points in the rejection region than will D_1, without increasing the overall probability above α.

(iii) It is consistent in the sense that if an observed event is significant at the $100\alpha_1$ % level, then it must also be significant at any $100\alpha_2$ % level, where $\alpha_2 > \alpha_1$.

(iv) There is a simple way of generalization to the case of multivariate discontinuous distributions, e.g. for the $h \times k$ contingency table.

It can, however, lead to certain anomalies which, as Prof. E. S. Pearson has pointed out to me, may not be regarded as acceptable. Consider the three examples given below.

The definition D_3 will be given later. In Example 1, D_2 includes $i = 0$ in the rejection region which D_1 does not, and this seems very reasonable. For the two-tailed 5 % level in Example 2, D_2 again gives a larger region than D_1, although it is open to question whether it would not have been better to include $i = 0$ rather than $i = 5$ in the region. When, however, we take a 7·2 % level* in this example, we find a more serious difficulty. Although

$$p(0) = 3{\cdot}53\ \% < \tfrac{1}{2} \times 7{\cdot}2\ \%,$$

D_2 does not make $i = 0$ significant, but chooses $i = 5$ instead. Thus in this instance D_2 contains no term or terms from the lower tail of the distribution.

* It should be made clear that anomalies of this kind arise only rarely and therefore we have illustrated this situation on Example 2, taking an unconventional significance level.

Example 1

a	*c*	*A*			33
b	*d*	*B*		*i*	7
R	*S*	*N*	=	25	15	40

If d is replaced by a variate i and the marginal totals are kept constant, the probability distribution of i (under the null hypothesis) is as follows:

i	0	1	2	3	4	5	6	7
$p(i)$ in %	2·58	14·25	29·92	30·87	16·84	4·83	0·67	0·04

Definition	Cases of i which are two-tailed significant at the 5 % level	
D_1		6, 7
D_2	0	6, 7
D_3	0	6, 7

Example 2

			33
		i	7
26	14		40

i	0	1	2	3	4	5	6	7
$p(i)$ in %	3·53	17·29	32·10	29·19	13·96	3·49	0·42	0·02

Definition	Significant events (values of i)	
	Two-tailed 5 % level	Two-tailed 7·2 % level
D_1	6, 7	0, 6, 7
D_2	5, 6, 7	5, 6, 7
D_3	0, 6, 7	0, 6, 7

Example 3

	i	31 12
31	12	43

i	0	1	2 3 4 5	6	7	8	9	10	11	12
$p(i)$ in %	0·920	6·624	...	4·435	0·877	0·102	0·006	0	0	0

Definition	Two-tailed significant at the 1 % level
D_1	8 9 10 11 12
D_2	7 8 9 10 11 12
D_3	8 9 10 11 12

If we consider what is the objection to the D_2 region in this last respect, we seem to reach the following broad conclusions. The purpose of a two-tailed test is to pick out departures from the null hypothesis in either direction, i.e. to establish significance when there is either a marked positive or negative association between X and Y. Clearly it can only act in this way if the rejection region contains terms from both tails of the distribution. In the case of extreme skewness where the first (or last) term has a probability that is greater than 100α %, 'two-tailedness' is obviously impossible, but in cases like those of Examples 2 and 3 it is *possible* to have a two-tailed 5 % (Ex. 2) and 1 % (Ex. 3) rejection region. What is less clear is how to define that region in general terms.

It is possible that a more fundamental attack on the problem could be based on a study of the power function of the test.* We shall, however, be content here to put forward a definition giving a region which must include that resulting from the customary D_1 but which, as with D_2, will often contain more points. If we start from the two-tailed region of D_1 there will often be only one choice of a point to be included in the wider region; this is the case for Example 1, where the 5 % rejection region under D_1 (including $i = 6, 7$) can only be extended by including $i = 0$. In the case of Example 2, however, the 5 % region of D_1 (6, 7) can be extended either by including $i = 0$ or 5 but not, of course, both.

A number of alternative definitions were considered; some of these were ambiguous or could lead to the inconsistency referred to under (iii) on p. 496 above. Finally, we have taken a definition D_3, which appears generally, though not quite always, to satisfy intuitive requirements for a two-tailed test.

Definition D_3. Define $F(E)$, or the 'first tail' probability, as the cumulative sum of the probabilities under the null hypothesis of all possible events which are more extreme, in the

* [In this connexion it is perhaps relevant to note that when considering, on the basis of the power function, an 'unbiased' two-tailed test in the case of continuous asymmetrical variation, Neyman & Pearson (*Statist. Res. Mem.* **2** (1936), 18–25) found that a larger probability (i.e. $> \frac{1}{2}\alpha$) should be cut off from the steeper tail of the null hypothesis distribution than from the finer tail.—Ed.]

same direction, than a given event E, including the probability of E itself. Define $S(E)$, or the 'second tail' probability, as the cumulative sum of probabilities, starting with that of the event most extreme in the opposite direction as compared with E, and cumulating up to but not exceeding the value of $F(E)$. If, and only if, $F(E)+S(E) \leqslant \alpha$, include E in the rejection region for the two-tailed 100α % level of significance.

The rejection regions defined by D_3 are shown below the table of probabilities for each of the Examples 1–3.

It can be readily seen that every event which is significant under D_1 will be significant under D_3, but the regions under D_2 and D_3 will not necessarily correspond. This last point is illustrated in Example 2, where D_3 may be thought to meet the requirements for a two-tailed best better than D_2. On the other hand, in the case of Example 3, where a two-tailed 1 % rejection region is required, D_3 gives a region no wider than D_1, failing to include the point $i = 7$. It is interesting to note that in this case there is a genuinely two-tailed region which could be used, namely, that including $i = 0, 9, 10, 11, 12$, giving a total probability of $0{\cdot}926 < 1{\cdot}000$ %. From the point of view of all-round power of discrimination it is possible that this last region might have some claim for consideration.

To illustrate anomalies and differences between definitions, we have naturally picked out exceptional cases, but it should be emphasized that these cases are very rare. Thus, among all the 2×2 tables with $N \leqslant 50$ considered in building up the tables of significance points given below, there were only thirty-two cases in which the regions defined by D_2 and D_3 differed (one case of disagreement usually results in two changesi n the tables).

In forming the tables with significance levels of 5 and 1 %, the computation was done independently three times, once on the basis of D_2 and later on the basis of D_3 and of another alternative definition, not given here, which was finally discarded for possible inconsistency. The tables contain the results according to D_3. The cases where the result of using D_2 is different are indicated by + or − signs above or below the corresponding entry; thus a + indicates that the entry is to be increased by 1, a − that it is to be decreased by 1, to obtain the limit under D_2.

In the following section we give directions for use of the tables with an example, while an Appendix contains the mathematical description of the method by which the tables with this special form of entry have been built up.

3. DIRECTIONS FOR USE OF THE TABLES

$N \leqslant 50$ is the number of observations.

The tables are sometimes applicable in cases $N > 50$. See Note 2.

No 2×2 table where $N < 6$ is significant one-tailed at the 5 % level.

Step 1. Arrange the given 2×2 table in the form

a	c	A
b	d	B
R	S	N

where A is the largest and B is the smallest (or in case of equality one of the smallest) of the four marginal totals, A, B, R, S, and $aB > Ab$.

If $aB = Ab$, the table is not significant.

Step 2. Note the four figures:

$$\left.\begin{array}{ll} d & \text{for finding the appropriate table,} \\ b & \\ c-b=S-B=A-R\geqslant 0 & \text{for finding the place in the table,} \\ a-d=R-B=A-S\geqslant 0 & \text{for decision.} \end{array}\right\}$$

Step 3. Look at the table at the end of this paper headed $i = d$ and find the intersection of arrays $x = b$ and $y = c-b$. There are four entries

$$\begin{matrix} z_1 & z_2 \\ z_3 & z_4 \end{matrix} \quad \text{(exceptions see Notes 1, 2 and 3)}$$

for the one-tailed and two-tailed points of significance according to the scheme:

	One-tailed	Two-tailed
5 %	z_1	z_2
1 %	z_3	z_4

Step 4. Compare the figure $a-d$ with the entry z_j, in which you are interested ($j = 1, 2, 3, 4$), e.g. the given 2×2 table is two-tailed significant at the 5 % level. if and only if $a-d \geqslant z_2$. Correspondingly for z_1, z_3 and z_4.

Note 1. If no entry is given, the answer is:

(*a*) Not significant for $N \leqslant 50$ if the missing entry should be to the right or below any given entry in the table headed $i = d$, or somewhere to the right of the whole table; e.g. $i = 11, x = 5, y = 14, a-d = 2$. The missing entry should be to the right of the given entry $z_1 = 8$, $i = 11$, $x = 3$, $y = 14$. Therefore the event $\begin{vmatrix} 13 & 19 \\ 5 & 11 \end{vmatrix}$ is not significant at the 5% level.

(*b*) Significant two-tailed at the 1 % level if the missing entry should be to the left or above any given entry in the table headed $i = d$ or somewhere to the left of the whole table; e.g. $i = 16$, $x = 2$, $y = 13$, $a-d = 1$. Because $x < 3$, the missing entry should be to the left of the whole table for $i = 16$. Therefore the event $\begin{vmatrix} 17 & 15 \\ 2 & 16 \end{vmatrix}$ is significant two-tailed at the top level.

(*c*) The entry '–' instead of a numerical entry stands for 'not significant for $N \leqslant 50$'.

Note 2. These tables are sometimes applicable in cases where $B+S \leqslant 50 < N$, i.e. in cases where an entry is given or missing to the left. These tables are not applicable if $50 < B+S$. In most practical cases (assuming $N > 50$) where these tables fail to give an answer, the Table VIII given by Fisher & Yates (1949) gives an answer and vice versa in all cases for $N \leqslant 50$.

Note 3. In cases where only two entries (counting '–' as an entry) are given instead of four, each entry stands for both the one-tailed and two-tailed point of significance.

Example for application of the tables

Given

4	18	22
12	9	21
16	27	43

Step 1

18	9	27
4	12	16
22	21	43

$B = 16 < 21 < 22 < 27,$

$A = 27 > 22 > 21 > 16,$

$a = 18, \quad b = 4,$

$Ba = 16 \times 18 = 288 > 108 = 27 \times 4 = Ab;$

consequently the arrangement is correct.

The only possible other arrangement after having chosen B as the smallest marginal total is

9	18	27
12	4	16
21	22	43

But now $a = 9, \quad b = 12,$

$Ba = 16 \times 9 = 144 < 324 = 27 \times 12 = Ab,$

contrary to the agreed kind of arrangement

Step 2

$d = 12$ for finding the appropriate table,

$b = x = 4$, $c - b = y = 5$ for finding the place in the table,

$a - d = 6$ for comparison with z_j.

Step 3. In the table headed $i = 12$ at the intersection of $x = 4$ and $y = 5$, the entries

$$\begin{matrix} z_1 & z_2 \\ z_3 & z_4 \end{matrix} \quad \text{are given as} \quad \begin{matrix} 1 & 2 \\ 6 & 8 \end{matrix}$$

Step 4. Result: the given 2×2 table is

significant (one-tailed) at the 5 % level $(6 > 1)$,
significant (two-tailed) at the 5 % level $(6 > 2)$,
significant (one-tailed) at the 1 % level $(6 = 6)$,
not significant (two-tailed) at the 1 % level $(6 < 8)$.

The entry $z_2 = 2$ shows that even the less significant table with the same d, B, S, (b and c), and with $a - d = 2$ or

14	9	23
4	12	16
18	21	39

is significant (two-tailed) at the 5 % level. But $z_1 = 1$ shows that the table with $a - d = 0$ or

12	9	21
4	12	16
16	21	37

is not significant (one-tailed) at the 5 % level.

$z_4 = 8$ shows that the first 2×2 table with the same d, B and S as the given table, which is significant (two-tailed) at the 1 % level, is with $a - d = 8$ or

20	9	29
4	12	16
24	21	45

Note 4. A user who wants to apply definition D_3 must neglect all + or − signs attached to any entry in the table.

A user who wants to apply definition D_2 must read a + sign attached to an entry as: add 1 to the given entry, and a − sign as: subtract 1 from the given entry. For example:

In the case $i = 5$, $x = 1$, $y = 12$, by $z_2 = \overline{21}$ is meant:

in case of D_3 $z_2 = 21$
in case of D_2 $z_2 = 21 - 1 = 20$

Correspondingly, in the case $i = 9$, $x = 0$, $y = 21$, by $z_4 = \overset{+}{9}$ is meant:

in case of D_3 $z_4 = 9$
in case of D_2 $z_4 = 9 + 1 = 10$

APPENDIX

Mathematical description of the tables

The tables are completely based on the well-known exact formula:

$$p\begin{pmatrix} a & c \\ b & d \end{pmatrix} = \frac{A!B!R!S!}{N!a!b!c!d!} = \frac{\binom{S}{d}\binom{R}{B-d}}{\binom{N}{B}} = p(d \mid BSN),$$

where $p(d \mid BSN)$ is the (single) probability of obtaining d with given marginal totals B and S and given number of observations N, assuming that both classifications are independent of each other.

It follows directly from the arrangement of the table [$B = \text{Min}(ABRS)$ and $Ba > Ab$] that $Nd > BS$ and $d \leqslant B \leqslant S \leqslant N - B$, in particular $2B \leqslant N$.

For given d, B, S and N the probability $p(i \mid BSN)$ ($i = 0, 1, 2, \ldots, B$) is a step function of i only. This function, call it $p(i)$, has the following properties:

It rises monotonically from $i = 0$ to a maximum which is at $i_{\max.} = [m]$, where

$$(N+2)\,m = (B+1)(S+1)$$

and falls monotonically until $i = B$.

There are some special cases:

(1) $i_{\max.}$ may be 0 or B, in which cases $p(i)$ is monotonically decreasing or increasing in its whole range.

(2) There are two equal values $p(i_{\max.}) = p(i_{\max.} - 1)$ in the cases where m is an integer.

Proof.

$$p(i) : p(i+1) = n : D, \text{ where } n = (i+1)(N-S-B+i+1) \text{ and } D = (B-i)(S-i).$$

Thus $p(i) \leqslant p(i+1)$ if $v(i) \leqslant 0$,
$p(i) \geqslant p(i+1)$ if $v(i) \geqslant 0$, where $v(i)(N+2) = n - D$, $v(i) = i + 1 - m$.

Therefore, if $i \geqslant i_{\max.}$ $v(i) \geqslant [m] + 1 - m > 0$,
if $i < i_{\max.}$ $v(i) \leqslant [m] - m \leqslant 0$,

which proves the above statements.

It may be shown that $$\left[\frac{BS}{N}\right] \leqslant i_{\max.} \leqslant \left[\frac{BS}{N}\right] + 1.$$

Definitions $f = f(d \mid BSN)$ or 'first tail':

$$f(d \mid BSN) = \sum_{i=d}^{B} p(i \mid BSN)$$

$s = s(d \mid BSN)$ or 'second tail':

$$s(d \mid BSN) = \sum_{i=0}^{k} p(i \mid BSN), \quad \text{where} \quad k < d,$$

$$p(k \mid BSN) \leqslant p(d \mid BSN), \quad p(k+1 \mid BSN) > p(d \mid BSN) \quad (k+1 = d \text{ in the case } d = i_{\max.}).$$

If no k exists (that is, if $p(0 \mid BSN) > p(d \mid BSN)$), then $s(d \mid BSN) = 0$.
[k is (given B and S) a function $k(d, N)$ and $Nk < BS$.]

$t = t(d \mid BSN)$ or 'two-tailed probability',
$t = f + s$.

From these definitions follows:

Statement 1. $p(i \mid BSN) > p(i \mid BS, N+1)$ if $i \geqslant d$
Statement 2. $f(d \mid BSN) > f(d \mid BS, N+1)$ } exactly.
Statement 3. $p(i \mid BSN) < p(i \mid BS, N+1)$ if $i \leqslant k(d, N)$ }
Statement 4. $t(d \mid BSN) \geqslant t(d \mid BS, N+1)$ *not* exactly, but with sufficient approximation in practice.

The construction of the tables is based on Statements 2 and 4.

For the one-tailed case, it would be enough to compute $f(d \mid BSN)$ for fixed d, B, S and increasing N until the table is found to be significant. Then Statement 2 enables the tables to be condensed to the given form. For the two-tailed case, Statement 4 could serve the same purpose if it were exact. The inexactness of Statement 4 was the reason for computing all possible cases until $N = 50$. The computation was done by direct evaluation of the exact formula for $t(d \mid BSN)$.

Thus the entries z_1 and z_3 give answers based on mathematically exact general reasons. The entries z_2 and z_4 are (although of course correct in the tables) based only on empirical computation. An exact proof of Statement 4 is impossible, as one can show examples where it is not true.

Proofs for Statements 1, 2 *and* 3

(1) $(h + BS) \cdot p(i \mid BSN) = [h + i(N+1)] \cdot p(i \mid BS, N+1)$, where $h = (N+1)(N+1-B-S)$. Statement 1 follows on account of $i(N+1) > iN \geqslant dN > BS$.

(2) Follows directly from 1.

(3) $$\frac{BS}{N} - \frac{BS}{N+1} = \frac{BS}{N(N+1)} \leqslant \frac{B(N-B)}{N(N+1)} \leqslant \frac{N}{4(N+1)} < \frac{1}{4}.$$

Corresponding to 1, one finds:

$$^{*}p(i, N) \geqslant p(i, N+1) \quad \text{as long as} \quad (N+1)\, i \geqslant BS,$$

and for $$(N+1)\, i < BS, \quad p(i, N) < p(i, N+1).$$

But for the single possible value i with $\dfrac{BS}{N+1} \leqslant i < \dfrac{BS}{N}$ follows $i = i_{\max.}$ because

$$m - \frac{BS}{N+1} = \frac{B+S+1}{N+2} - \frac{BS}{(N+1)(N+2)} < 1,$$

and this, together with $i \leqslant k(d, N)$, contradicts the definition of k which proves Statement 3.

Conclusions from Statement 3

(1) $k(d, N)$ is a monotonically decreasing step function of N, or $k(d, N) \geqslant k(d, N+1)$.

(2) As long as $p[k(d, N) \mid BS, N+1] \leqslant p(d \mid BS, N+1)$, it follows that $s(d \mid BSN) < s(d \mid BS, N+1)$, but in the cases where $p[k(d, N) \mid BS, N+1] > p(d \mid BS, N+1)$, there is one member less in $s(N+1)$ than in $s(N)$, and thus $s(d \mid BSN) > s(d \mid BS, N+1)$.

(It is easy to prove that $p(i \mid BSN) > p(i-1 \mid BS, N+1)$ if $i \leqslant k(d, N)$.)

* By $p(i, N)$ is meant $p(i \mid BSN)$.

Thus the second tail as a function of N increases monotonically step by step in intervals of N, where

$$k(d, N+1) = k(d, N),$$

and decreases in big steps from N to $N+1$, where

$$k(d, N+1) < k(d, N)$$

(or where the length of the second tail decreases).

Of course, $\lim_{N\to\infty} s(d \mid BSN) = 0$, because $\lim_{N\to\infty} i_{\max.} = 0$.

Question of Statement 4

$\lim_{N\to\infty} t(d \mid BSN) = 0$, because $t = s+f$, $\lim_{N\to\infty} s = 0$, and $\lim_{N\to\infty} f = 0$, the last equation following from:

$$\lim_{N\to\infty} i_{\max.} = 0 \quad \text{and} \quad \lim_{N\to\infty} p(0, N) = \lim_{N\to\infty} \frac{\binom{N-S}{B}}{\binom{N}{B}} = 1.$$

The question now is: Is this decreasing of $t(d \mid BSN)$ with increasing N monotonic like that of $f(d \mid BSN)$ or not? Unfortunately it is not.

Proof. Computation gives, e.g. for

$$d = 9, \quad B = 9, \quad S = 36, \quad N = 48,$$
$$t(9 \mid 9, 36, 48) < 0{\cdot}0883,$$
$$t(9 \mid 9, 36, 49) > 0{\cdot}0892.$$

Therefore, Statement 4 is certainly not exact.*

The size of $t(N)$ and $t(N+1)$ is influenced mainly by the sums

$$\tau_N = p(d \mid BSN) + p(k \mid BSN) \quad \text{and} \quad \tau_{N+1} = p(d \mid BS, N+1) + p(k \mid BS, N+1),$$

where both times $k = k(d, N)$.†

The second term in τ_{N+1} is included in $t(N+1)$ only under the condition

$$p(d, N+1) \geqslant p(k, N+1).$$

Thus only these cases must be considered.

We have

$$\tau_N = \frac{\phi_N + \sigma_N}{D_N},$$

where

$$\phi_N = \binom{S}{d}\binom{N-S}{B-d}, \quad \sigma_N = \binom{S}{k}\binom{N-S}{B-k}, \quad D_N = \binom{N}{B},$$

$k = k(d, N)$ in σ_{N+1} also.

It is easy to show by simple computing that

$$\tau_N . D_{N+1}\left(1 - \frac{B}{N+1}\right) = \phi_{N+1}\left(1 - \frac{B-d}{N+1-S}\right) + \sigma_{N+1}\left(1 - \frac{B-k}{N+1-S}\right),$$

and the ratio τ_N/τ_{N+1} will be smallest (that is, the 'bad' case for Statement 4) if

$$\phi_{N+1} = \sigma_{N+1} \quad \text{or} \quad \tau_{N+1} D_{N+1} = 2\phi_{N+1},$$

and in this worst case one has

$$\tau_N\left(1 - \frac{B}{N+1}\right) = \tau_{N+1}\left\{1 - \frac{2B-(d+k)}{2(N+1-S)}\right\}.$$

Thus, in the worst case $\tau_N < \tau_{N+1}$ is equivalent to

$$\frac{B}{N+1} < \frac{2B-(d+k)}{2(N+1-S)} \quad \text{or} \quad 2BS > (N+1)(d+k).$$

* This difficulty does not arise if the definition D_2 is replaced by D_1, because then Statement 4 is exact. But in case of D_3, this example applies and D_3 does not improve the situation as compared with D_2.

† If k does not exist, then $s = 0$ and Statement 4 is exact.

Now both the members in this 'bad' inequality are usually of roughly the same size, because:

$$\text{if } 2S \leqslant N+1 \text{ then } d+k \leqslant B,$$
$$\text{if } 2S \geqslant N \text{ then } d+k \geqslant B,$$

and in the case $2S = N+1$ both members are equal and therefore

$$t(d \mid B, S, 2S) = t(d \mid B, S, 2S-1),$$

and only in the rare cases where the three conditions:

(1) $\phi_{N+1} \approx \sigma_{N+1}$,
(2) $\phi_{N+1} \geqslant \sigma_{N+1}$,
(3) $2BS > (d+k)(N+1)$,

are all given is it possible to have $\tau_{N+1} > \tau_N$,

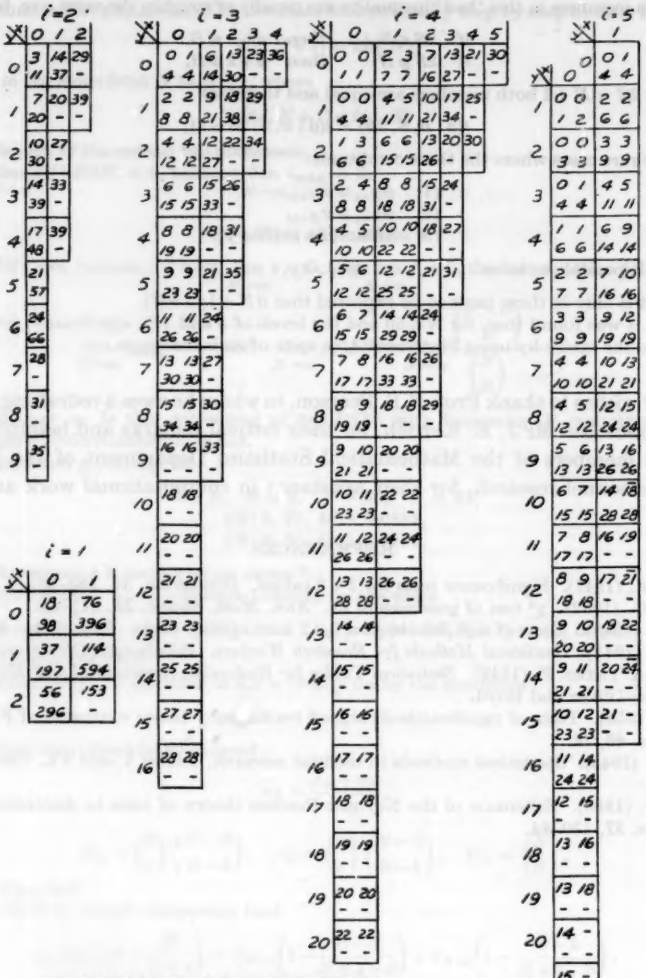
and therefore it is only in these cases to be expected that $t(N+1) > t(N)$.

Empirically, it was found that, for $N \leqslant 50$ and the levels of 5 and 1 % significance, no inconsistency is introduced to the tables by using Statement 4, in spite of its inexactness.

The author wishes to thank Prof. E. S. Pearson, to whom he owes a redrafting of §§ 1 and 2, Mr A. G. Arbous and Mr J. E. Kerrich, for their critical remarks and helpful suggestions, and also the members of the Mathematical Statistics Department of the National Institute for Personnel research, for their assistance in computational work and editing of this paper.

REFERENCES

BARNARD, G. A. (1947). Significance tests for 2×2 tables. *Biometrika*, **34**, 123–38.
COCHRAN, W. G. (1952). χ^2 test of goodness of fit. *Ann. Math. Statist.* **23**, 315–45.
FINNEY, D. J. (1948). Tests of significance in a 2×2 contingency table. *Biometrika*, **35**, 148.
FISHER, R. A. (1941). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.
FISHER, R. A. & YATES, F. (1949). *Statistical Tables for Biological, Agricultural and Medical Research*. Edinburgh: Oliver and Boyd.
LATSCHA, R. (1953). Tests of significance in a 2×2 contingency table: extension of Finney's tables. *Biometrika*, **40**, 74.
MAINLAND, D. (1948). Statistical methods in medical research. Tables V and VI. *Canad. J. Res.* E, **26**, 1–116.
TOCHER, K. D. (1950). Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika*, **37**, 130–44.



i = 1

i \ j	0	1
0	18 98	76 396
1	37 197	114 594
2	56 296	153 -

i = 2

i \ j	0	1	2
0	3 11	14 37	29 -
1	7 20	20 -	39 -
2	10 30	27 -	
3	14 39	33 -	
4	17 48	39 -	
5	21 57		
6	24 66		
7	28 -		
8	31 -		
9	35 -		

i = 3

i \ j	0	1	2	3	4
0	0 1 4 4	6 14	13 30	23 -	36 -
1	2 2 8 8	9 21	18 38	29 -	
2	4 4 12 12	12 27	22 -	34 -	
3	6 6 15 15	15 33	26 -		
4	8 8 19 19	18 -	31 -		
5	9 9 23 23	21 -	35 -		
6	11 11 26 26	24 -			
7	13 13 30 30	27 -			
8	15 15 34 34	30 -			
9	16 16 - -	33 -			
10	18 18 - -				
11	20 20 - -				
12	21 21 - -				
13	23 23 - -				
14	25 25 - -				
15	27 27 - -				
16	28 28 - -				

i = 4

i \ j	0	1	2	3	4	5
0	0 0 1 1	2 3 7 7	7 16	13 27	21 -	30 -
1	0 0 4 4	4 5 11 11	10 21	17 34	25 -	
2	1 3 6 6	6 6 15 15	13 26	20 -	30 -	
3	2 4 8 8	8 8 18 18	15 31	24 -		
4	4 5 10 10	10 10 22 22	18 -	27 -		
5	5 6 12 12	12 12 25 25	21 -	31 -		
6	6 7 15 15	14 14 29 29	24 -			
7	7 8 17 17	16 16 33 33	26 -			
8	8 9 19 19	18 18 - -	29 -			
9	9 10 21 21	20 20 - -				
10	10 11 23 23	22 22 - -				
11	11 12 26 26	24 24 - -				
12	13 13 28 28	26 26 - -				
13	14 14 - -					
14	15 15 - -					
15	16 16 - -					
16	17 17 - -					
17	18 18 - -					
18	19 19 - -					
19	20 20 - -					
20	22 22 - -					

i = 5

i \ j	0	1
0	0 1 4 4	
1	0 0 1 2	2 2 6 6
2	0 0 3 3	3 3 9 9
3	0 1 4 4	4 5 11 11
4	1 1 6 6	6 9 14 14
5	2 2 7 7	7 10 16 16
6	3 3 9 9	9 12 19 19
7	4 4 10 10	10 13 21 21
8	4 5 12 12	12 15 24 24
9	5 6 13 13	13 16 26 26
10	6 7 15 15	14 18 28 28
11	7 8 17 17	16 19 - -
12	8 9 18 18	17 21 - -
13	9 10 20 20	19 22 - -
14	9 11 21 21	20 24 - -
15	10 12 23 23	21 - - -
16	11 14 24 24	
17	12 15 - -	
18	13 16 - -	
19	13 18 - -	
20	14 - - -	
21	15 - - -	
22	16 - - -	
23	17 - - -	

ARRANGEMENT OF THE ENTRIES

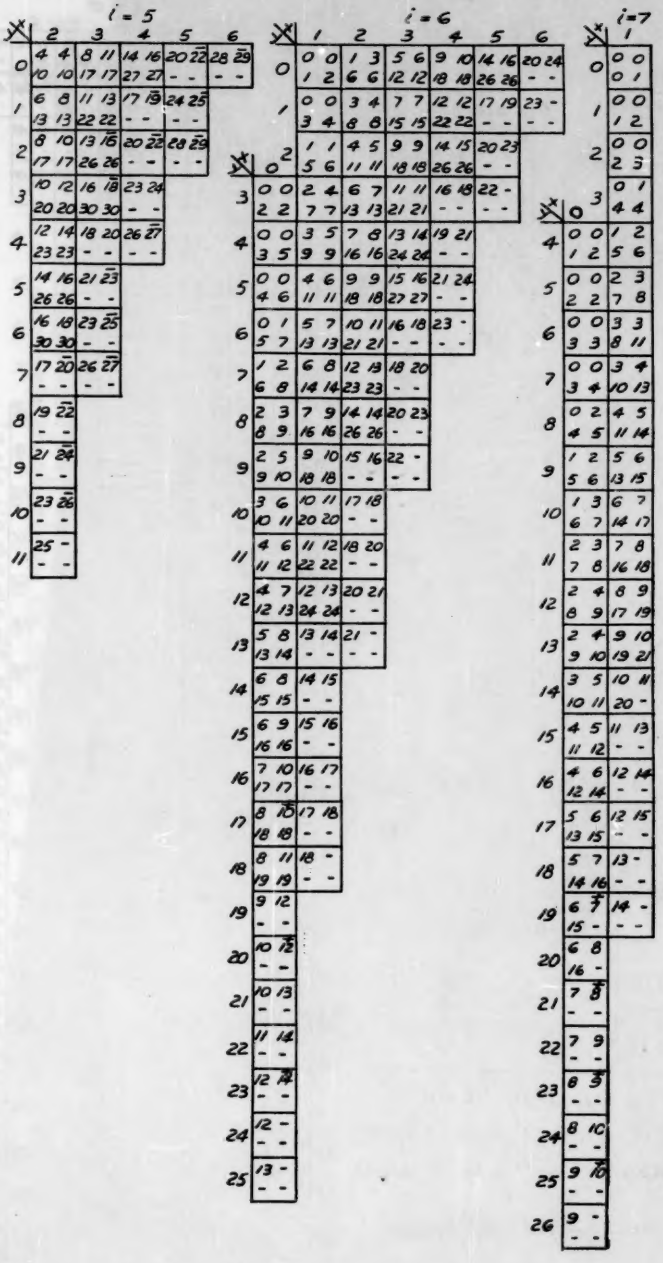
	ONE-TAILED	TWO-TAILED
5%	z_1	z_2
1%	z_3	z_4

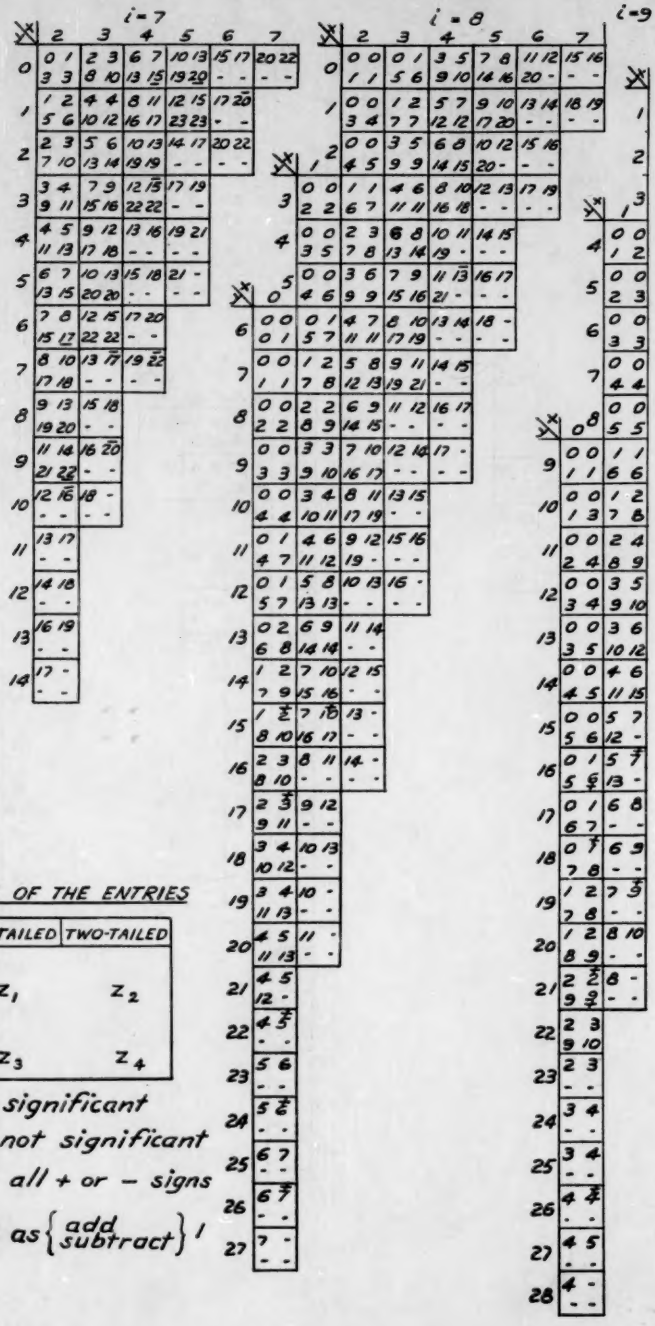
$a - d \geq z_j$, significant

$< z_j$, not significant

D_3: neglect all + or − signs

D_2: read $\mp$ as $\left\{\begin{matrix}\text{add}\\\text{subtract}\end{matrix}\right\}$ 1


i = 5
i = 6
i = 7



ARRANGEMENT OF THE ENTRIES

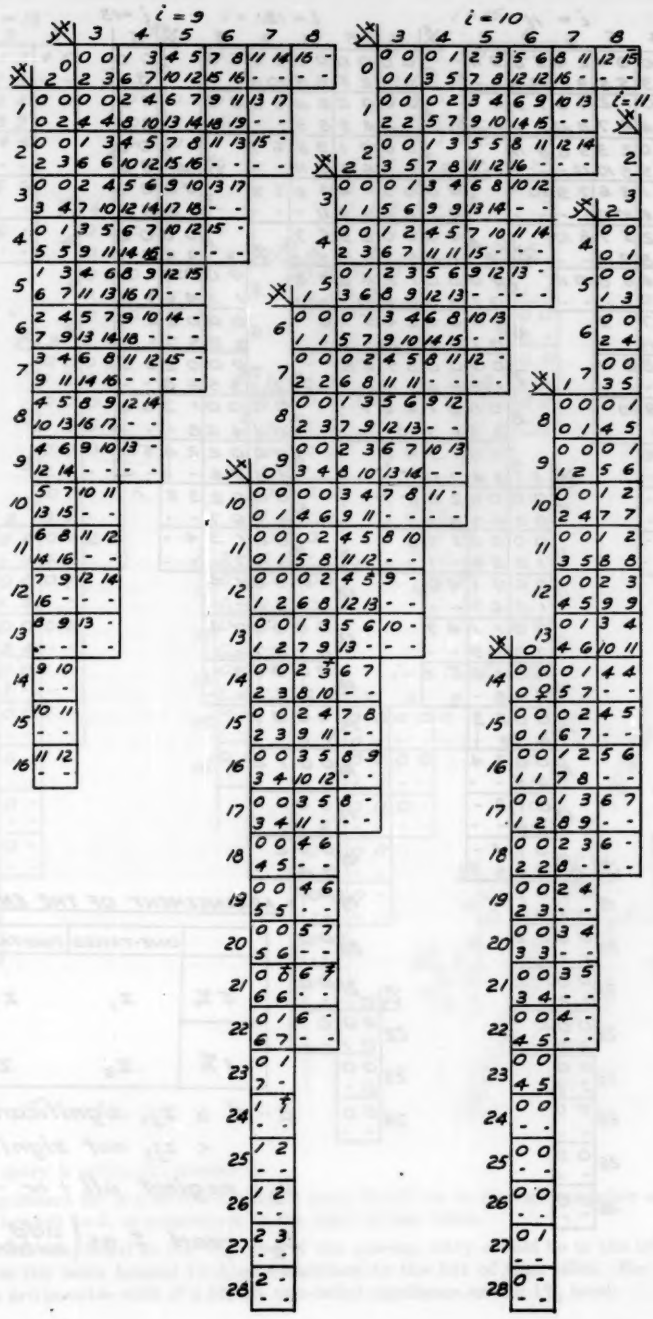
	ONE-TAILED	TWO-TAILED
5 %	z_1	z_2
1 %	z_3	z_4

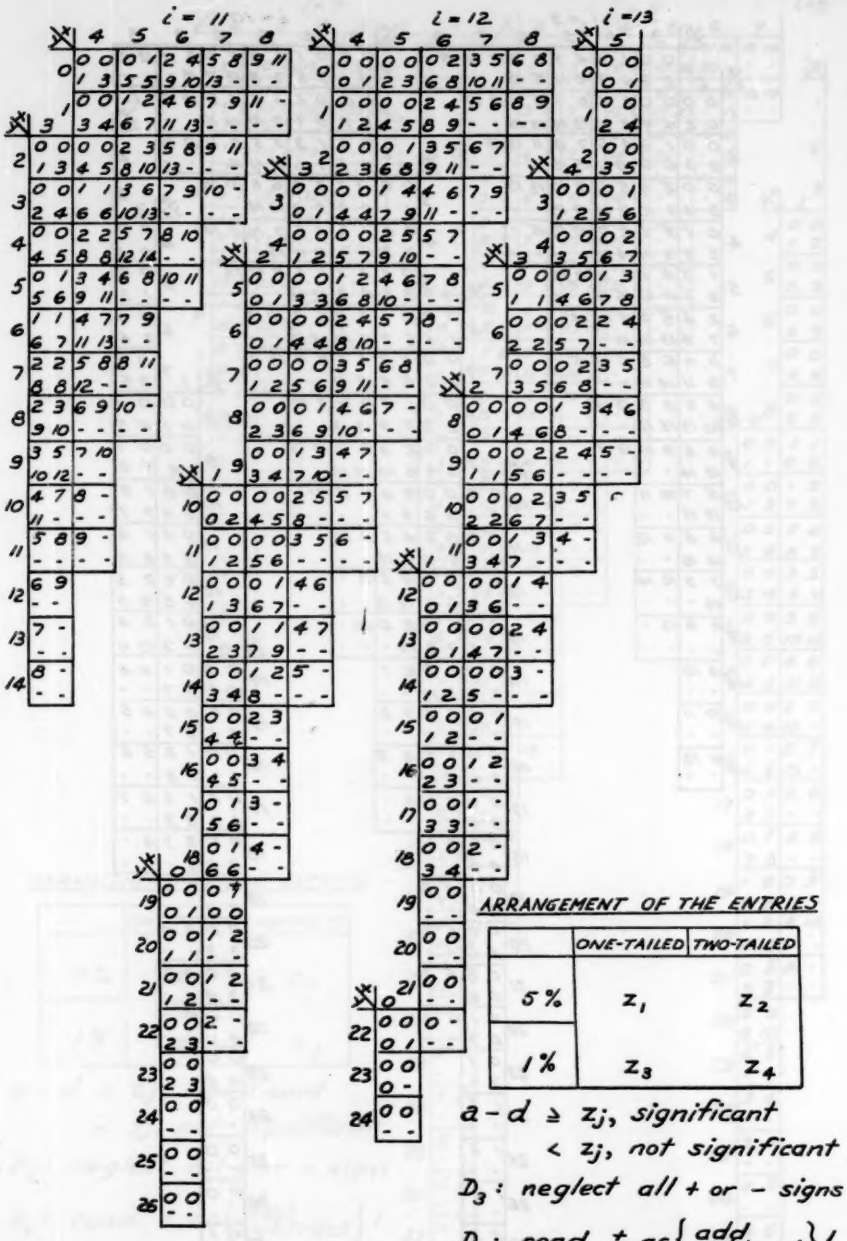
$a - d \geq z_j$, significant

$< z_j$, not significant

D_3: neglect all + or − signs

D_2: read ± as $\left\{\begin{matrix}\text{add}\\ \text{subtract}\end{matrix}\right\}$ 1


i = 9
i = 10
i = 11



ARRANGEMENT OF THE ENTRIES

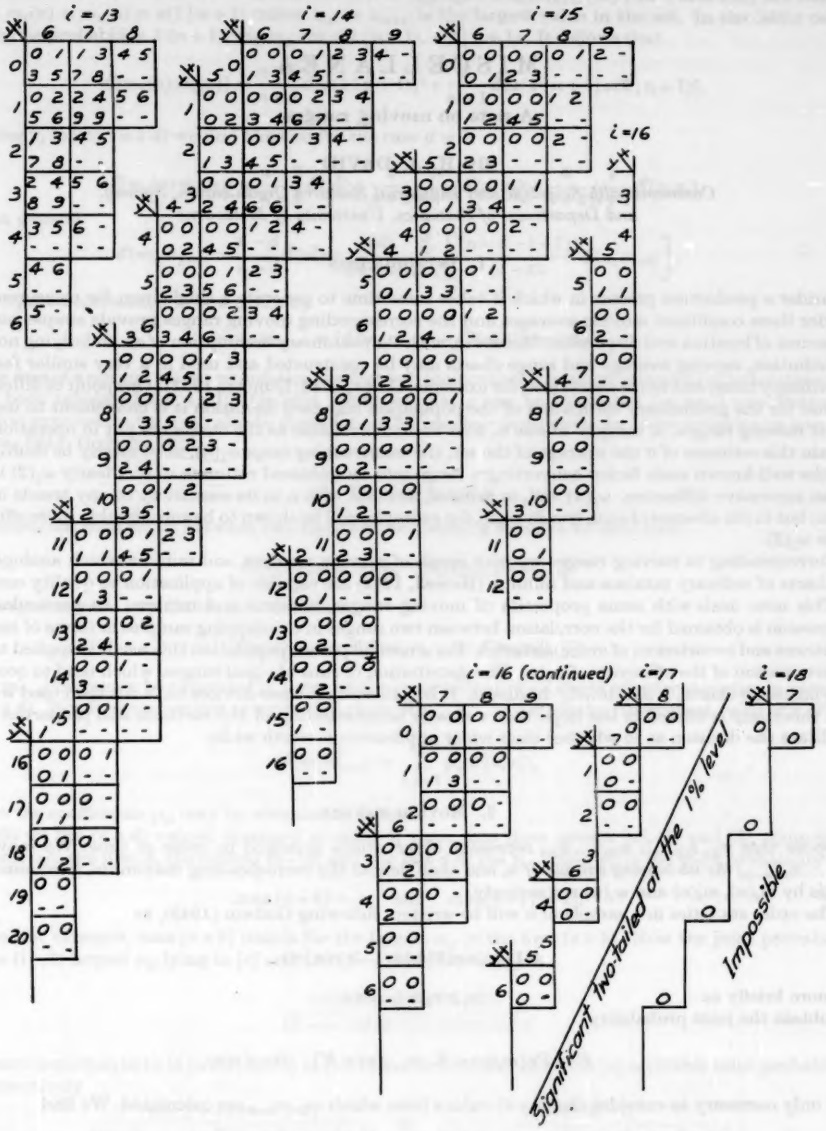
	ONE-TAILED	TWO-TAILED
5%	z_1	z_2
1%	z_3	z_4

$a - d \geq z_j$, significant

$< z_j$, not significant

D_3: neglect all + or − signs

D_2: read ± as {add / subtract} 1



Note: If no entry is given, the answer is:

(*a*) Not significant for $N \leqslant 50$ if the missing entry should be to the right or below any given entry in the table headed $i=d$, or somewhere to the right of that table.

(*b*) Significant two tailed at the 1% level if the missing entry should be to the left or above any given entry in the table headed $i=d$, or somewhere to the left of that table. For $i \geqslant 18$ all 2×2 tables which are possible with $N \leqslant 50$, are two-tailed significant at the 1% level.

MISCELLANEA

A note on moving ranges

By H. A. DAVID

Commonwealth Scientific and Industrial Research Organization, Sydney, and Department of Statistics, University of Melbourne

1. Introduction

Consider a production process in which it takes some time to generate a single item for measurement. Under these conditions moving averages and the corresponding moving ranges provide simple current measures of location and dispersion. Moreover, with the customary assumption of an underlying normal distribution, moving average and range charts may be constructed and used in a very similar fashion to ordinary mean and range charts (see, for example, Grant, 1946; Duncan, 1952). One point of difference is that for the preliminary estimation of the population standard deviation it is convenient to use the set of moving ranges, in samples of size n, that become available as the charts are put in operation. To obtain this estimate of σ the average of the set, the *mean moving range* $\omega_1(n)$, need merely be multiplied by the well-known scale factor converting a range into an unbiased estimate of σ. Clearly $\omega_1(2)$ is the mean successive difference. $\omega_1(n)$ will, in general, increase with n in its sensitivity to any trends in the data; but in the absence of such trends $\omega_1(5)$, for example, will be shown to be considerably more efficient than $\omega_1(2)$.

Corresponding to moving ranges we may speak of moving maxima and minima which analogously to charts of ordinary maxima and minima (Howell, 1949) are capable of application to quality control.

This note deals with some properties of moving ranges, maxima and minima. In particular, an expression is obtained for the correlation between two ranges in overlapping samples in terms of means, variances and covariances of order statistics. For a normal parent population this result is applied to the determination of the efficiency of $\omega_1(n)$. The distribution of runs of equal ranges, which tend to occur in moving range charts, is also briefly discussed. It is realized that these devices have not been used widely and this study is offered in the hope that a clearer understanding of the methods and properties may facilitate the decision as to whether their wider application is worth while.

2. Moving maxima

Suppose that $x_{(1)}, x_{(2)}, \ldots, x_{(i)}, \ldots, x_{(N)}$ represent observations arranged in order of time. We may call $(x_{(i)} \ldots x_{(i+n-1)})$ *the ith moving sample of* n, and shall denote the corresponding maximum, minimum and range by $m_i(n)$, $m_i'(n)$ and $w_i(n)$ respectively.

The order statistics in a sample of n will be written, following Godwin (1949), as

$$x(1\,|\,n) \geqslant x(2\,|\,n) \geqslant \ldots \geqslant x(n\,|\,n);$$

or more briefly as

$$x_1 \geqslant x_2 \geqslant \ldots \geqslant x_n.$$

To obtain the joint probability

$$P = \Pr[m_i(n) < X,\ m_{i+d}(n) < Y] \quad (0 < d < n)$$

it is only necessary to consider the $(n+d)$ values from which m_i, m_{i+d} are calculated. We find

$$P = F^n(X)\,F^d(Y) \quad (X \leqslant Y),$$

$$= F^d(X)\,F^n(Y) \quad (X \geqslant Y),$$

where

$$F(X) = \Pr(x < X).$$

However, we proceed to determine the correlation between m_i and m_{i+d} by a less direct method. The x's will henceforth be assumed continuous.

Consider first two successive maxima and the set $x_{(1)}, x_{(2)}, \ldots, x_{(n)}, x_{(n+1)}$ (say) on which they are based. Then $m_1(n) = m_2(n) = x(1 \mid n+1)$ unless $x_{(1)}$ or $x_{(n+1)}$ is the largest value in the set. In the latter case, which has probability $2/(n+1)$, m_1, m_2 are $x(1 \mid n+1)$, $x(2 \mid n+1)$. It follows that

$$\mathscr{E}[m_1(n)\, m_2(n)] = \frac{n-1}{n+1}\mathscr{E}[x(1 \mid n+1)]^2 + \frac{2}{n+1}\mathscr{E}[x(1 \mid n+1)\, x(2 \mid n+1)].$$

Writing x_i for $x(i \mid n+d)$ we find likewise for the case $d = 2$

$$\mathscr{E}[m_1(n)\, m_3(n)] = \frac{n-2}{n+2}\mathscr{E}(x_1^2) + \frac{4}{n+2}\frac{n}{n+1}\mathscr{E}(x_1 x_2) + \frac{4}{n+2}\frac{1}{n+1}\mathscr{E}(x_1 x_3),$$

and in general

$$\mathscr{E}(m_1 m_{1+d}) = \frac{n-d}{n+d}\mathscr{E}(x_1^2) + \frac{2d!}{(n+d)!}\sum_{t=1}^{d}\left[\frac{(n-d-t-1)!}{(d-t)!}\mathscr{E}(x_1 x_{1+t})\right].$$

The required correlation is now given by

$$\rho(m_1, m_{1+d}) = [\mathscr{E}(m_1 m_{1+d}) - \mathscr{E}^2(m_1)]/\mathrm{var}\,(m_1).$$

In the case of a normal parent population the first two moments and product-moments of order statistics have been tabulated for $n \leqslant 10$ (Godwin, 1949), so that ρ can be calculated for $n \leqslant 5$ (see Table 1). Explicit expressions for ρ can also be given for $n \leqslant 3$ from the exact results for order statistics up to $n = 6$ (Jones, 1948; Godwin, 1949).

3. Moving ranges

To obtain the correlation between two ranges in overlapping samples we note that

$$\mathscr{E}(w_1 w_{1+d}) = \mathscr{E}(m_1 - m_1')\,(m_{1+d} - m_{1+d}').$$

For a symmetrical population this reduces to

$$\mathscr{E}(w_1 w_{1+d}) = 2[\mathscr{E}(m_1 m_{1+d}) - \mathscr{E}(m_1 m_{1+d}')].$$

$\mathscr{E}(m_1 m_{1+d}')$ can again be expressed in terms of the expectations of products of order statistics in samples of $(n+d)$. It will be convenient to write x_j' in place of $x_{n+d+1-j}$ to denote the jth smallest value of x. Then

$$\mathscr{E}(m_1 m_{1+d}') = \sum_{i,j=1}^{1+d} p_{ij}\mathscr{E}(x_i x_j'),$$

where the coefficients p_{ij} may be obtained as follows.

Split up the $(n+d)$ values, arranged in order of time, into three groups $[a]$, $[b]$ and $[c]$, comprising respectively the first d, the central $(n-d)$ and the last d of the $x_{(i)}$. Then p_{ij} is the joint probability of

$$\max[a+b] = x_i \quad \text{and} \quad \min[b+c] = x_j',$$

where, for example, $\max[a+b]$ stands for the largest $x_{(i)}$ in the first $(a+b)$. Now the joint probability of the $(i-1)$ largest $x_{(i)}$ lying in $[c]$ and the $(j-1)$ smallest in $[a]$ is

$$\frac{(d!)^2\,(n+d-i-j+2)!}{(d-i+1)!\,(d-j+1)!\,(n+d)!}.$$

We next require x_i to lie in $[a+b]$ and x_j' in $[b+c]$. According as x_i falls into $[a]$ or $[b]$ this joint probability is respectively

$$\frac{(d-j+1)\,(n-i+1)}{(n+d-i-j+2)\,(n+d-i-j+1)} \quad \text{or} \quad \frac{(n-d)\,(n-i)}{(n+d-i-j+2)\,(n+d-i-j+1)}.$$

Hence we have

$$p_{ij} = \frac{(d!)^2\,(n+d-i-j)!}{(d-i+1)!\,(d-j+1)!\,(n+d)!}[(d-j+1)\,(n-i+1) + (n-d)\,(n-i)].$$

The required correlation coefficients are now readily obtained and, for a normal parent, are given in Table 1 up to $n = 5$.

Table 1. *Correlations between maxima and ranges in overlapping samples of n having (n − d) values in common*

n	d	Maxima	Ranges
2	1	0·4264	0·2239
3	1	0·5990	0·5395
	2	0·2743	0·2062
4	1	0·6923	0·6651
	2	0·4309	0·3942
	3	0·2028	0·1732
5	1	0·7506	0·7353
	2	0·5322	0·5097
	3	0·3374	0·3149
	4	0·1611	0·1456

4. The efficiency of mean moving ranges

When moving ranges are employed in quality control it is convenient to use as the preliminary estimate of σ, calculated from N observations, the statistic

$$\omega_1 = \sum_{i=1}^{N-n+1} w_i/(N-n+1).$$

More generally, a mean moving range ω_r may be constructed based on every rth sample.

The efficiency E of ω_r defined by

$$E = \operatorname{var} s'/\operatorname{var} \omega_r,$$

where

$$s' = \Gamma\left(\frac{N-1}{2}\right)\sqrt{\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]}\Big/\left[\sqrt{2}\,\Gamma\left(\frac{N}{2}\right)\right],$$

is given in Table 2 in a number of cases with $n \leqslant 5$. For purposes of comparison we list E also for the ordinary mean range estimator $\bar{w}$ in samples of five and the most efficient (non-overlapping) weighted mean range estimator w^* (see Grubbs & Weaver, 1947).

Table 2. *Efficiency of various mean range estimators based on samples of n from N observations*

N	ω_1				ω_2	$\bar{w}$	w^*
	$n=2$	$n=3$	$n=4$	$n=5$	$n=4$	$n=5$	—
10	0·643	0·735	0·753	0·747	0·785	0·826	0·850
20	0·623	0·730	0·767	0·772	0·772	0·773	0·799
30	0·617	0·729	0·775	0·788	0·769	0·756	0·786
40	0·614	0·729	0·778	0·796	0·768	0·748	0·778
50	0·612	0·729	0·780	0·801	0·768	0·743	0·773
60	0·611	0·729	0·782	0·804	0·767	0·740	0·769
80	0·609	0·729	0·784	0·808	0·767	0·736	0·767
100	0·608	0·729	0·785	0·811	0·766	0·734	0·762
200	0·607	0·729	0·788	0·816	0·766	0·729	0·759
∞	0·605	0·728	0·790	0·821	0·765	0·725	0·754

It will be noted that E is lowest for ω_1 with $n=2$. This is, of course, the case of the mean successive difference which is, however, less influenced by trends than the statistics based on larger n.

Compared with $\overline{w}$ and w^* the ω-statistics are seen to gain in efficiency as N increases. Their relatively low efficiency for small N is presumably due to the uneven weighting of the observations which is inherent in their construction.

For fixed n all the statistics ω, $\overline{w}$ and w^* tend to normality with increasing N. This is obvious for $\overline{w}$ and w^* and has, in fact, been proved explicitly for ω_1 (Hoel, 1946).

5. Runs of equal ranges

Two ranges in samples of n, with $(n-d)$ observations in common, will be equal if two of these $(n-d)$ are the extremes in the $(n+d)$ values involved. Thus *for any continuous population* the probability of equality is

$$P_{d+1} = [(n-d)(n-d-1)]/[(n+d)(n+d-1)].$$

Clearly, P_{d+1} may be interpreted as the probability of a run of $(d+1)$ or more equal ranges, so that the distribution of runs $p(r)$ is determined. It is easily shown that $p(r)$ is J-shaped, flattening out with increasing n.

The expected number of runs is given by

$$\mathscr{E}(r) = \sum_{d=0}^{n-2} P_{d+1} = (3n-2) - 2(2n-1)[\psi(2n-1) - \psi(n)],$$

where $\psi(n) = d\log\Gamma(n)/dn$. $\mathscr{E}(r)$ is evaluated in Table 3 for $n \leqslant 20$ together with the expected number of runs of equal maxima, viz.

$$\sum_{d=0}^{n-1} (n-d)/(n+d) = n[2\psi(2n) - 2\psi(n) - 1].$$

Table 3. *Expected number of runs of equal maxima and ranges*

n	Maxima	Ranges	n	Maxima	Ranges
2	1·333	1·000	8	3·606	2·239
3	1·700	1·167	9	3·991	2·462
4	2·076	1·367	10	4·375	2·687
5	2·456	1·579	12	5·146	3·137
6	2·839	1·796	15	6·303	3·815
7	3·222	2·017	20	8·232	4·947

REFERENCES

Duncan, A. J. (1952). *Quality Control and Industrial Statistics.* Chicago: Richard D. Irwin.
Godwin, H. J. (1949). *Ann. Math. Statist.* **20**, 279.
Grant, E. L. (1946). *Statistical Quality Control.* New York and London: McGraw-Hill.
Grubbs, F. E. & Weaver, L. C. (1947). *J. Amer. Statist. Ass.* **42**, 224.
Hoel, P. G. (1946). *Ann. Math. Statist.* **17**, 475.
Howell, J. M. (1949). *Ann. Math. Statist.* **20**, 305.
Jones, H. L. (1948). *Ann. Math. Statist.* **19**, 270.

Censored samples from truncated normal distributions*

BY A. CLIFFORD COHEN, JR.
The University of Georgia

1. INTRODUCTION AND SUMMARY

In life-testing and response-time studies, selection procedures sometimes operate to effect a truncation at the lower end of the time scale prior to starting a test which is subsequently terminated before all sample specimens exhibit the reaction being studied. Resulting samples may thus be regarded as censored from a truncated population. The present paper is limited to censored samples from truncated normal distributions. It is related to previous studies of truncated and censored samples by Hald (1949), the author (1950, 1954), Gupta (1952), and various other writers. It is also related to a paper on life testing by Epstein & Sobel (1953), in which some of the advantages of employing censored samples to conserve time and test specimens are discussed with regard to a one-parameter exponential distribution. For samples of the types considered here, maximum-likelihood estimators (estimates) of the population mean and standard deviation are derived, and their asymptotic variances are obtained. An illustrative example is given to demonstrate the practical application of these results.

2. MAXIMUM-LIKELIHOOD ESTIMATION

The probability density (frequency) function of a normal distribution with mean, m, and standard deviation, σ, truncated on the left at a fixed terminus, x_0, may be written as

$$\left.\begin{aligned} f(x) &= (I_1\sigma\sqrt{(2\pi)})^{-1}\exp-[(x-m)^2/(2\sigma^2)] \quad (x_0\leqslant x\leqslant\infty),\\ f(x) &= 0 \qquad\qquad\qquad\qquad\qquad\quad (x<x_0), \end{aligned}\right\} \tag{1}$$

where $I_1 = I(\xi_1)$, is the proportion of the complete distribution retained after truncation and ξ_1 is the left terminus (truncation point) in standard units of the complete distribution, i.e.

$$\xi_1 = (x_0-m)/\sigma, \quad I(\xi) = \int_{\xi}^{\infty}\phi(t)\,dt, \quad \phi(t) = (2\pi)^{-\frac{1}{2}}\exp(-\tfrac{1}{2}t^2). \tag{2}$$

Let N sample specimens be selected from a population distributed according to (1) and let their life spans or reaction times $\{x_i\}$ $(i = 1, 2, 3, \ldots, N)$ be observed and recorded until a fixed number $(n<N)$ have reacted. Let observation be discontinued upon determining x_n. Thus x_n is the greatest of the measured observations, and it is known that each of the $N-n$ censored observations exceeds that value. When x is distributed according to (1), the likelihood function for a sample of the type described is

$$P = \left\{K(\sigma I_1\sqrt{(2\pi)})^{-n}\exp\left[-\sum_1^n(x_i-m)^2/(2\sigma^2)\right]\right\}(I_2/I_1)^{N-n}, \tag{3}$$

where K is a constant,

$$\xi_2 = (x_n-m)/\sigma \quad \text{and} \quad I_2 = I(\xi_2). \tag{4}$$

According to Gupta (1952), a sample in which the number of censored (unmeasured) observations is fixed, is designated as Type II censored as distinguished from Type I censored samples in which the terminal is fixed. If rather than stopping a test after n (fixed) observations have been made, it is terminated at the expiration of a fixed time, x_u, then n is a random variable and the resulting sample is of Type I. The likelihood function of a Type I censored sample differs from (3) only in that K is replaced by a different constant, and in (4), x_n is replaced by x_u. Consequently, the same estimators are applicable in both cases except for the trivial case of a Type I censored sample in which $n = 0$. In this latter situation, the estimators would not be defined.

Taking logarithms of (3), differentiating and equating to zero, we have

$$\left.\begin{aligned} \frac{\partial L}{\partial m} &= -\frac{N}{\sigma}\frac{\phi_1}{I_1}+\frac{1}{\sigma}\sum_1^n\left(\frac{x_i-m}{\sigma}\right)+\frac{N-n}{\sigma}\frac{\phi_2}{I_2} = 0,\\ \frac{\partial L}{\partial\sigma} &= -\frac{N}{\sigma}\frac{\xi_1\phi_1}{I_1}-\frac{n}{\sigma}+\frac{1}{\sigma}\sum_1^n\left(\frac{x_i-m}{\sigma}\right)^2+\frac{N-n}{\sigma}\frac{\xi_2\phi_2}{I_2} = 0, \end{aligned}\right\} \tag{5}$$

where $L = \log P$ and $\phi_i = \phi(\xi_i)$ $(i = 1, 2)$.

* Sponsored by the Office of Ordnance Research, U.S. Army.

Let Z_i designate the reciprocal of Mill's ratio, i.e.

$$Z_i = \phi_i/I_i = e^{-\frac{1}{2}\xi_i^2} \Big/ \int_{\xi_i}^{\infty} e^{-\frac{1}{2}t^2}\,dt \quad (i = 1, 2). \tag{6}$$

Substitute (6) into (5) and simultaneously substitute $m = x_0 - \sigma\xi_1$, which follows from (2). After simplifying, we obtain the estimating equations

$$\left.\begin{aligned} \nu_1/w &= \{(N/n)\,Z_1 - [(N-n)/n]\,Z_2 - \xi_1\}/(\xi_2 - \xi_1), && (a)\\ \bar{s}^2/w^2 &= \{1 + (N/n)\,\xi_1 Z_1 - [(N-n)/n]\,\xi_2 Z_2 - [(N/n)\,Z_1 - ((N-n)/n)\,Z_2]^2\}/(\xi_2 - \xi_1)^2, && (b) \end{aligned}\right\} \tag{7}$$

where w is the restricted sample range, ν_k is the kth sample moment about the left terminus and $\bar{s}^2$ is the sample variance, i.e.

$$w = x_n - x_0, \quad \nu_k = \sum_1^n (x_i - x_0)^k/n, \quad \bar{s}^2 = \nu_2 - \nu_1^2. \tag{8}$$

The two equations of (7) can be solved simultaneously for $\hat{\xi}_1$ and $\hat{\xi}_2$ as illustrated in § 4. With these estimates determined, estimates of the mean and standard deviation follow from (2) and (4) as

$$\hat{\sigma} = w/(\hat{\xi}_2 - \hat{\xi}_1) \quad \text{and} \quad \hat{m} = x_0 - \hat{\sigma}\hat{\xi}_1. \tag{9}$$

The standard maximum-likelihood symbol ($\hat{}$) serves to distinguish estimates from parameters estimated.

Estimating equations (7) can be expressed in a form that is analogous to corresponding equations previously given by the author (1950) for doubly truncated normal samples. However, Z_i as used here is not defined quite the same as in the earlier paper. Here $Z_i = Z(\xi_i)$ is the reciprocal of Mill's ratio and is a function of ξ_i only, whereas for the doubly truncated cases, Z_1 and Z_2 were each defined as functions of both ξ_1 and ξ_2.

When truncation is on the right and censoring is on the left, it is merely necessary to reverse the signs of all observations and proceed as for the case discussed above since with $f(x)$ truncated on the right, $f(-x)$ will be truncated on the left.

3. Variances of estimates

The variance-covariance matrix of $(\hat{m}, \hat{\sigma})$ is derived from the second-order partial derivatives of L. We let $\phi_{11}(\xi_1, \xi_2)$, $\phi_{12}(\xi_1, \xi_2)$ and $\phi_{22}(\xi_1, \xi_2)$ designate stochastic limits of $-\frac{\sigma^2}{n}\frac{\partial^2 L}{\partial m^2}$, $-\frac{\sigma^2}{n}\frac{\partial^2 L}{\partial m\,\partial\sigma}$ and $-\frac{\sigma^2}{n}\frac{\partial^2 L}{\partial\sigma^2}$ as $n \to \infty$. Thus we have

$$\left.\begin{aligned} \phi_{11}(\xi_1, \xi_2) &= 1 - (N/n)\,Z_1(Z_1 - \xi_1) + [(N-n)/n]\,Z_2(Z_2 - \xi_2),\\ \phi_{12}(\xi_1, \xi_2) &= (N/n)\,Z_1[1 - \xi_1(Z_1 - \xi_1)] - [(N-n)/n]\,Z_2[1 - \xi_2(Z_2 - \xi_2)],\\ \phi_{22}(\xi_1, \xi_2) &= 2 + (N/n)\,\xi_1 Z_1[1 - \xi_1(Z_1 - \xi_1)] - [(N-n)/n]\,\xi_2 Z_2[1 - \xi_2(Z_2 - \xi_2)]. \end{aligned}\right\} \tag{10}$$

The asymptotic variances and the covariance are then given as

$$\left.\begin{aligned} \operatorname{var}(\hat{m}) &\sim [\hat{\sigma}^2/n]\,[\hat{\phi}_{22}/(\hat{\phi}_{11}\hat{\phi}_{22} - \hat{\phi}_{12}^2)],\\ \operatorname{var}(\hat{\sigma}) &\sim [\hat{\sigma}^2/n]\,[\hat{\phi}_{11}/(\hat{\phi}_{11}\hat{\phi}_{22} - \hat{\phi}_{12}^2)],\\ \operatorname{cov}(\hat{m}, \hat{\sigma}) &\sim [\hat{\sigma}^2/n]\,[-\hat{\phi}_{12}/(\hat{\phi}_{11}\hat{\phi}_{22} - \hat{\phi}_{12}^2)], \end{aligned}\right\} \tag{11}$$

where $\hat{\phi}_{ij}$ is written for $\phi_{ij}(\hat{\xi}_1, \hat{\xi}_2)$.

It subsequently follows that the correlation between $\hat{m}$ and $\hat{\sigma}$ is given by

$$\rho_{\hat{m},\hat{\sigma}} \sim -\hat{\phi}_{12}/\sqrt{(\hat{\phi}_{11}\hat{\phi}_{22})}. \tag{12}$$

As given above, (10), (11) and (12) are for the Type II censored case in which N and n are fixed while x_n is a random variable. These same results also apply in the Type I case in which N and x_u are fixed while n is a random variable, if we replace n, N/n and $(N-n)/n$ by their expected values,

$$E(n) = N(I_1 - I_2)/I_1, \quad E(N/n) = I_1/(I_1 - I_2) \quad \text{and} \quad E[(N-n)/n] = I_2/(I_1 - I_2).$$

4. An illustrative example

One hundred units of a certain type of electronic device are selected for life testing from a group which has survived 600 hr. of prior service. The number of units which expired prior to the time of selection is unknown. The total life in hours is recorded as each of the first ninety units of the sample expires, and

the test is terminated as soon as the ninetieth unit expires. For purposes of this illustration, we consider logarithms $\{x\}$ of the life spans to be normally distributed so that for the sample selected, we have:

$$N = 100, \quad n = 90, \quad x_0 = \log 600 = 2{\cdot}778151, \quad w = 0{\cdot}235276, \quad \nu_1 = 0{\cdot}12560147,$$
$$\bar{s}^2 = 0{\cdot}00348667813, \quad \nu_1/w = 0{\cdot}533847354 \quad \text{and} \quad \bar{s}^2/w^2 = 0{\cdot}062987824.$$

To solve estimating equations (7), we employ an iterative procedure with initial approximations read from a chart recently given by the writer (1954) for graphically solving estimating equations of doubly truncated samples. By this procedure, information provided by knowledge of the number of censored observations is neglected until subsequent iterations. Thereby we obtain $\xi_1^{(0)} = -1{\cdot}60$ and $\xi_2^{(0)} = 1{\cdot}21$. To improve these initial approximations, we substitute $\xi_1^{(0)} = -1{\cdot}600$ into (7*b*) and by inverse interpolation obtain $\xi_2^{(1)} = 1{\cdot}302$. This value is then substituted into (7*a*) and we find $\xi_1^{(1)} = -1{\cdot}663$. Repeating the cycle, we find $\xi_2^{(2)} = 1{\cdot}287$ and $\xi_1^{(2)} = -1{\cdot}626$. In the notation employed here $\xi_i^{(j)}$ is the jth approximation to ξ_i. Closer approximations to the required estimates could be reached by continuing the iterations described above through additional cycles. However, greater improvement can be achieved with less labour if we take advantage of the fact that the iterants already obtained locate two points, $P_1(\xi_1^{(0)}, \xi_2^{(1)})$ and $P_3(\xi_1^{(1)}, \xi_2^{(2)})$, which lie on the curve defined by (7*b*), and two points, $P_2(\xi_1^{(1)}, \xi_2^{(1)})$ and $P_4(\xi_1^{(2)}, \xi_2^{(2)})$, which lie on the curve defined by (7*a*). We approximate the two curves in this vicinity by straight lines through these two pairs of points, and coordinates of their intersection provide improved approximations to the required estimates.

Since P_1 and P_2 each have the same ordinate, and since P_3 and P_4 likewise have the same ordinate, the required coordinates are readily determined by interpolation as summarized below.

ξ_2	ξ_1 by (7*a*)	ξ_1 by (7*b*)	Diff.
1·302	−1·663	−1·600	−0·063
1·293	−1·640	−1·640	0
1·287	−1·626	−1·663	+0·037

Thus we have $\hat{\xi}_1 = -1{\cdot}640$ and $\hat{\xi}_2 = 1{\cdot}293$, and from (9) it follows that

$$\hat{\sigma}_x = 0{\cdot}080217, \quad \hat{m}_x = 2{\cdot}90971.$$

Since $x = \log T$, where T is the life in hours, we estimate the mean life in hours as $\hat{m}_T = 812{\cdot}3$ hr.

If additional decimal places had been required in the above results, we could have continued the iterative process described above through one or more further cycles.

Variances and covariances of the above estimates may be computed using (10) and (11). Although ordinary tables of normal curve areas and ordinates are adequate for this purpose, tables provided by Sampford (1952) greatly facilitate the computations. Sampford tabulated

$$\lambda = Z(Z - \xi) \quad \text{and} \quad \zeta = Z[1 - \xi(Z - \xi)],$$

with η designating the argument rather than ξ as here. In using his tables, however, a word of caution is necessary since an unfortunate printing error resulted in negative signs before some of the entries of ζ whereas all of these entries should be positive.

With $\hat{\xi}_1 = -1{\cdot}640$ and $\hat{\xi}_2 = 1{\cdot}293$, we interpolate from Sampford's tables, and using (10) compute $\hat{\phi}_{11} = 0{\cdot}87961$, $\hat{\phi}_{12} = 0{\cdot}39418$ and $\hat{\phi}_{22} = 1{\cdot}12908$. On substituting these values in (11) and (12), we compute

$$\text{var}(\hat{m}) \sim 0{\cdot}0000964, \quad \text{var}(\hat{\sigma}) \sim 0{\cdot}0000751 \quad \text{and} \quad \rho_{\hat{m},\hat{\sigma}} \sim -0{\cdot}3955.$$

REFERENCES

Cohen, A. C., Jr. (1950). Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. *Ann. Math. Statist.* **21**, 557.

Cohen, A. C., Jr. (1954). Truncated and censored samples. Tech. Rep. no. 11, Ordnance research contract DA-01-009-ORD-288. *Univ. Ga. Dep. Math.*

Epstein, B. & Sobel, M. (1953). Life testing. *J. Amer. Statist. Ass.* **48**, 486.

GUPTA, A. K. (1952). Estimation of the mean and standard deviation of a normal population from a censored sample. *Biometrika*, **39**, 260.

HALD, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Skand. AktuarTidskr.* **32**, 119.

SAMPFORD, M. R. (1952). The estimation of response-time distributions. II. *Biometrics*, **8**, 307.

The rapid calculation of χ^2 as a test of homogeneity from a $2 \times n$ table

BY J. B. S. HALDANE

Department of Biometry, University College, London

In genetical work we very frequently have to test a $(2 \times n)$-fold table for homogeneity. We count, for example, non-recombinants and recombinants in an experiment on linkage. The frequency of the latter is unknown or only roughly known before the experiment. Only if the values found in several families are homogeneous have we the right to assume that the value derived from the total has the precision given by its standard sampling error. Suppose our table to be

a_1 b_1	a_2 b_2	—	a_r b_r	—	a_n b_n	A B
s_1	s_2	—	s_r	—	s_n	N

That is to say the rth sample of s_r members consists of a_r members of one class and b_r of the other, while $\Sigma a_r = A$, $\Sigma b_r = B$, $A + B = N$. The exact contribution of the rth sample to χ^2 as a test of homogeneity is $\dfrac{(a_r B - b_r A)^2}{AB s_r}$, or one of many equivalent expressions.

Suppose $A > B$. If k is the nearest integer to AB^{-1}, $(A - kB)/N$ may be small, say $< 0{\cdot}1$. If not we can take a convergent k/h to A/B, where h and k are small integers, and $(hA - kB)/N$ will be small.

Then the following is an exact expression for the 'homogeneity χ^2' with $n-1$ degrees of freedom, where h may be unity:

$$\chi^2_h = \frac{N^2}{(h+k)^2 AB}\left[\sum_{r=1}^{n} (ha_r - kb_r)^2 s_r^{-1} - (hA - kB)^2 N^{-1}\right], \tag{1}$$

whilst the following is in error by a factor of less than $\left[1 - \dfrac{(hA - kB)}{N}\right]$:

$$\chi^2_h = \frac{1}{hk}\left[\sum_{r=1}^{n} (ha_r - kb_r)^2 s_r^{-1} - (hA - kB)^2 N^{-1}\right]. \tag{2}$$

If A and B are of the order of 1000, $(Ba_r - Ab_r)^2$ will be an integer with about five more figures than $(ha_r - kb_r)^2$, and the saving of labour is very considerable, while for smaller values most of the calculations can be reduced to mental arithmetic. The proof is as follows.

Let $\chi^2_r = \dfrac{(Ba_r - Ab_r)^2}{ABs_r}$, so $\chi^2_h = \sum_{r=1}^{n} \chi^2_r$.

Let $x_r = a_r - As_r N^{-1} = Bs_r N^{-1} - b_r$.

Then $\Sigma x_r = 0$ and $\chi^2_r = N^2 A^{-1} B^{-1} x_r^2 s_r^{-1}$.

Also $ha_r - kb_r = (hA - kB)\, s_r N^{-1} + (h+k)\, x_r$.

So $(ha_r - kb_r)^2 s_r^{-1} = (hA - kB)^2 N^{-2} s_r + 2(h+k)(hA - kB)\, x_r + (h+k)^2 x_r^2$

$= (hA - kB)^2 N^{-2} s_r + 2(h+k)(hA - kB)\, x_r + (h+k)^2 ABN^{-2}\chi^2_r$.

Hence $\sum_{r=1}^{n} [(ha_r - kb_r)^2 s_r^{-1}] = (hA - kB)^2 N^{-1} + (h+k)^2 ABN^{-2}\chi^2_h$.

whence (1) follows. But if $hA - kB = D$, then

$$A = \frac{kN+D}{h+k}, \quad B = \frac{hN-D}{h+k}, \quad \frac{N^2}{(h+k)^2 AB} = (hk)^{-1}\left[1 + \frac{(h-k)D}{hkN} - \frac{D^2}{hkN^2}\right]^{-1},$$

whence (2) follows.

As an example I take Table 1, where a_r and b_r are the numbers of normal and *vestigial Drosophila melanogaster* counted in eleven bottles in a class experiment where a 3 : 1 ratio was expected on Mendelian theory, but *vestigial* was known to be somewhat inviable. The question arose whether the mortality of recessives had been significantly different in different bottles. Clearly $h = 1$, $k = 6$, gives a good fit.

Table 1

a_r	b_r	s_r	$a_r - 6b_r$	$(a_r - 6b_r)^2/s_r$
25	1	26	+19	13·885
80	15	95	−10	1·053
38	12	50	−34	23·120
52	8	60	+ 4	0·267
9	0	9	+ 9	9·000
21	7	28	−21	15·750
33	6	39	− 3	0·231
24	2	26	+12	5·538
30	7	37	−12	3·892
51	7	58	+ 9	1·397
56	3	59	+38	24·475
419	68	487	+11	98·608

Here $h = 1$, $k = 6$, $A = 419$, $B = 68$, $N = 487$, $\sum_{r=1}^{n} (a_r - 6b_r)^2 s_r^{-1} - (A - 6B)^2 N^{-1} = 98{\cdot}360$. So formula (2) gives $\chi^2 = 16{\cdot}393$, formula (1) $\chi^2 = 16{\cdot}709$. There are 10 degrees of freedom, so $0{\cdot}09 > P > 0{\cdot}08$, and the data are not significantly heterogeneous. Had the expression $\Sigma[(Ba_r - Ab_r)^2 s_r^{-1}]/(AB)$ been used, the first summand would have been $1281^2/26$. In practice only two decimal places were used in the sum, and formula (2) was used, the whole calculation being completed on the blackboard in about ten minutes.

It is possible to justify formula (2) verbally as follows. $\Sigma[(h_r a_r - kb_r)^2 n_r^{-1}]$ is the value of χ^2 for n degrees of freedom, if we assume $\mathscr{E}(a_r) = ks_r/(h+k)$; while $(hA - kB)^2 N^{-1}$ is the value of χ^2 for A and B on the same assumption. Subtracting this we obtain χ_h^2 with $n-1$ degrees of freedom. In fact, the dissection of a χ^2 into components is only valid when all samples are large, and in this case leads to an error of 2%. While this dissection is often justifiable, it is desirable to know the magnitude of its error.

The 'Inefficiency' of the sample median for many familiar symmetric distributions*

By JOHN T. CHU, *University of North Carolina*

1. A lower bound for the variance of the median

If the reciprocal of the (asymptotic) variance of an estimate is taken as a measure of its (asymptotic) efficiency, the sample median $\tilde{x}$ is often (asymptotically) less efficient than the sample mean $\bar{x}$, for many symmetric distributions familiar to statisticians. In fact, for a symmetric distribution having its absolute maximum at the point of symmetry, if $\tilde{x}$ is asymptotically less efficient than $\bar{x}$, then quite often $\tilde{x}$ is never so efficient as $\bar{x}$, with the possible exception of very small samples. To show these facts, we first derive a very simple, yet sharp, lower bound for the variance of the sample median.

* Sponsored by the Office of Naval Research.

Suppose that $F(x)$ and $f(x)$ are the cumulative distribution function and the probability density function of a certain continuous distribution, and $f(x)$ is symmetric with respect to $x = \xi$, and $f(\xi) \geqslant f(x)$ for all x. Let $\tilde{x}$ be the median of a sample of size $2n+1$; then, writing $C_n = (2n+1)!/(n!\,n!)$,

$$\begin{aligned}\operatorname{var}\tilde{x} &= \int_{-\infty}^{\infty} (x-\xi)^2 C_n [F(x)]^n [1-F(x)]^n f(x)\,dx \\ &= \int_0^1 (x-\xi)^2 C_n F^n (1-F)^n\,dF \\ &\geqq [f(\xi)]^{-2} \int_0^1 (F-\tfrac{1}{2})^2 C_n F^n (1-F)^n\,dF \\ &= \{4[f(\xi)]^2 (2n+3)\}^{-1}. \qquad (1)\end{aligned}$$

The equality holds for a rectangular distribution.

2. Examples

It is well known that for normal and rectangular distributions $\bar{x}$ is more efficient than $\tilde{x}$. We shall show that this is true for many other familiar symmetric distributions.

(2·1) *Triangular distribution.* $f(x) = 1 - |x|$, $|x| \leqslant 1$. The variance of $\bar{x}$ is $\frac{1}{6}(2n+1)^{-1}$ and from (1) it follows that $\tilde{x}$ is less efficient than $\bar{x}$ for samples of sizes $2n+1$ provided $n > 1$. Direct computation shows that this is also true if $n = 1$.

(2·2) *t-Distribution.* $f(t) = A(1+t^2/k)^{-\frac{1}{2}(k+1)}$, where $A^{-1} = \Gamma(\frac{1}{2}k)\,(k\pi)^{\frac{1}{2}}/\Gamma[\frac{1}{2}(k+1)]$. Since $\sigma_t^2 = k/(k-2)$, it follows that for a t-distribution with k degrees of freedom, $\tilde{x}$ is less efficient than $\bar{x}$ if (but not necessarily only if)

$$\frac{\pi(k-2)\,\Gamma^2(\frac{1}{2}k)}{4\Gamma^2(\frac{1}{2}k+\frac{1}{2})} > \frac{2n+3}{2n+1}.$$

For both $k = 2m$ and $k = 2m+1$, the left-hand side of the above inequality is an increasing function of m. Computation shows, e.g. that the inequality holds if $k \geqslant 5$ and $n \geqslant 25$.

(2·3) *Symmetric β-distribution.*

$$f(x) = \frac{\Gamma(2p)}{\Gamma^2(p)} x^{p-1}(1-x)^{p-1} \quad (0 < x < 1;\ p > 0).$$

For this distribution $\sigma_x^2 = \frac{1}{4}(2p+1)^{-1}$. Hence $\tilde{x}$ is less efficient than $\bar{x}$ if

$$2^{4p-4}(2p+1)\,\Gamma^4(p)/\Gamma^2(2p) > \frac{2n+3}{2n+1}.$$

The left-hand side becomes smaller if p is replaced by $p+1$, and tends to $\frac{1}{2}\pi$ as $p \to \infty$. So it has a lower bound $\frac{1}{2}\pi$. Hence the inequality holds for every $p > 0$ and $n \geqslant 2$.

(2·4) *Cauchy type distribution.* This is defined to be of the type $f(x) = C_\alpha/(1+|x|^\alpha)$ $(-\infty < x < \infty;\ \alpha > 1)$. If $\alpha = 2$, we obtain the well-known Cauchy distribution for which $\tilde{x}$ is infinitely more efficient than $\bar{x}$. It would be interesting to examine whether or not $\bar{x}$ becomes more efficient as α increases. Now $\bar{x}$ has a finite variance only if $\alpha > 3$. C_α and $\operatorname{var}\bar{x}$ can be obtained by using contour integration (see Whittaker & Watson, 1952, p. 118). It follows that $\tilde{x}$ is less efficient than $\bar{x}$ if

$$\frac{y^2 \sin 3y}{\sin^3 y} > \frac{2n+3}{2n+1},$$

where $y = \pi/\alpha$. The left-hand side is a decreasing function of y, and so an increasing function of α. The least values of α for which the left-hand side is equal to 5/3 and 1 (the maximum and mi nimum of the right-hand side), are found to be 4·65 and 3·75 approximately.

3. Remarks

(i) $\bar{x}$ is not more efficient than $\tilde{x}$ for all symmetric distributions. When the parent population has a Laplace distribution, for example, $\tilde{x}$ is more efficient for all samples of odd sizes (Chu & Hotelling).

(ii) If $f(x)$ satisfies certain continuity conditions, $\tilde{x}$ has an asymptotically normal distribution and the asymptotic variance is $\{4[f(\xi)]^2(2n+1)\}^{-1}$. If the sample size, therefore, is not too small, the asymptotic variance is for all practical purposes a lower bound for the variance of $\tilde{x}$. And if $\tilde{x}$ is asymptotically less efficient than $\bar{x}$, then $\tilde{x}$ is less efficient than $\bar{x}$ for all samples whose sizes are not too small.

REFERENCES

Chu, J. T. & Hotelling, H. The moments of the sample median (to be published).

Whittaker, E. T. & Watson, G. N. (1952). *Modern Analysis.* Cambridge University Press.

A simple method of calculating the exact probability in 2 × 2 contingency tables with small marginal totals

By P. H. LESLIE

Bureau of Animal Population, Department of Zoological Field Studies, Oxford

It frequently happens that in the case of data which may be arranged in the form of a 2×2 contingency table, we wish to determine the exact probability of obtaining the observed result. In the method which appears to be adopted usually, we have to calculate the value of a number of expressions involving factorials. Although the actual calculations are not difficult, they become somewhat tedious even with the help of a table of log factorials, and a simple method of obtaining the exact solution may therefore be of practical use. The following method is convenient when the marginal totals are small (and usually it is only in such cases that we need to calculate the exact solution), and only a table of the binomial coefficients up to, say, $n = 20$ is required. Tables of these coefficients seem to be published up to $n = 12$, e.g. Barlow, and Milne-Thomson & Comrie, four-figure tables; but they may be extended once and for all to higher values very simply by means of Pascal's triangle.

Suppose we write any 2×2 table in the standard form:

x		n_A
		$N-n_A$
n_B	$N-n_B$	N

where N is the total number of observations, and the marginal totals fulfil the following inequalities,

$$n_A \leqslant N-n_A, \quad n_B \leqslant N-n_B \quad \text{and} \quad n_B \leqslant n_A;$$

so that n_B is the smallest marginal total, and x the number of observations in the cell with the smallest expectation. Then the set of possible results which is compatible with the given marginal totals is obtained by giving x in turn the successive values $0, 1, 2, \ldots, n_B$, and the probability of obtaining a particular value of x is given by the appropriate term of the hypergeometric distribution,

$$P(x) = \binom{n_A}{x}\binom{N-n_A}{n_B-x} \Big/ \binom{N}{n_B}.$$

If, as a first step, we write down opposite the values of $x = 0, 1, 2, \ldots, n_B$, starting at $x = 0$, the successive binomial coefficients for $n = n_A$, which we will call column a_x; and if, secondly, starting at the bottom opposite $x = n_B$, we write down in reverse order the binomial coefficients for $n = N-n_A$, and call this column b_x; then the product column $a_x b_x$ gives the successive terms of the distribution we require. The total of this last column,

$$\sum_x a_x b_x = \frac{N!}{n_B!(N-n_B)!},$$

which serves as a check if it is needed. The calculation of the exact probability $\sum_0^x P(x)$ or $\sum_x^{n_B} P(x)$ (whichever tail of the distribution is appropriate in the particular case), then follows very quickly on an ordinary calculating machine; in fact, the individual product terms, $a_x b_x$, need not necessarily be written down.

Merely as an arithmetical example, suppose we had the following table, where $n_A = 9$, $n_B = 8$, $N = 20$ and $x = 7$,

7	2	9
1	10	11
8	12	20

Then the two essential columns are the binomial coefficients for $n = 9$, forming column a_x, and those for $n = 11$, forming column b_x. Thus

x	a_x	b_x
0	1	165
1	9	330
2	36	462
3	84	462
4	126	330
5	126	165
6	84	55
7	36	11
8	9	1

Then, the sum of the product column, $\Sigma a_x b_x = 125{,}970$, which can be obtained directly on the machine without writing down the individual terms. Since the observed number, $x = 7$, in the cell with the smallest expectation is greater than expectation, we would require in this case the right-hand tail of the distribution. Thus

$$(36 \times 11) + (9 \times 1) = 405,$$

and

$$\sum_{7}^{8} P(x) = 405/125970 = 0{\cdot}003215.$$

A test for a change in a parameter occurring at an unknown point

By E. S. PAGE

Department of Mathematics, University of Durham

1. Introduction and summary

Consider a sample of independent observations in the order in which they were obtained, $x_1 \ldots x_n$; it is sometimes required to test the null hypothesis that all the observations are drawn from the same population with distribution function $F(x \mid \theta)$ against the alternative that $x_1, \ldots, x_m$ come from $F(x \mid \theta)$, and $x_{m+1}, \ldots, x_n$ from $F(x \mid \theta')$ $(\theta' \neq \theta)$. If m is known this is a straightforward problem of comparing two samples. In this paper we suppose that m is unknown; this raises new problems. A test is proposed for a case where θ is known and some comments are made on the problems presented by other cases.

2. One-sided case: θ known

Suppose that the initial value, θ, is known, One possibility for a test is to regard all the observations as a single sample and to use the best test that all the observations are from $F(x \mid \theta)$ against the alternative that all are from $F(x \mid \theta')$ for some $\theta' \neq \theta$. Such a test cannot be expected to be very powerful if the change occurs late in the sample; the few observations on the new parameter value would be obscured by the many on the old parameter. Since the problem of detecting a change in a parameter is important in controlling the quality of the output from a continuous production process, it is reasonable to investigate whether the methods of process inspection schemes can provide useful tests for the case in which we are interested.

A process inspection scheme for detecting a change in one direction in the parameter was given by the author in an earlier paper (Page, 1954). We suppose throughout that the parameter under consideration is the mean of the distribution unless the contrary is stated, and in this section we further assume that the value, θ, at the start of the observations is known. The scheme consisted of recording the cumulative sums $S_r = \sum_{i=1}^{r} (x_i - \theta)$, $S_0 = 0$, and taking action to rectify a possible change in the parameter when $S_r - \min_{0 \leqslant i < r} S_i \geqslant h$, i.e. when the sample path rises a height h above its previous minimum value; this

procedure can be displayed clearly on a chart. If there is no change in θ, the mean path of the cumulative sum is horizontal, while if an increase in θ occurs the new mean path has positive slope so that the above criterion would be satisfied without too much delay. The significance test suggested by this procedure is as follows:

I. *Given the observations* $x_1 \dots x_n$. *It is required to test the hypothesis that the mean is constantly* θ. *Use as the test statistic* $m = \max_{0 \leqslant r \leqslant n} \{S_r - \min_{0 \leqslant i < r} S_i\}$, *where* $S_r = \sum_{j=1}^{r} (x_j - \theta)$, $S_0 = 0$, *taking large values as significant, i.e. reject the hypothesis if* $m \geqslant h$.

It was shown in the paper cited that the properties of the corresponding process inspection scheme depended upon the characteristics of linear sequential tests; as the test I is a truncated form of the process inspection scheme it is to be expected that in general the properties of the test will be difficult to evaluate. A special case that is tractable is where the observations are nought-or-one binomial variables. Accordingly, we consider a test for the general case using binomial variables.

II. *Given the observations* $x_1 \dots x_n$. *It is required to test the hypothesis that the mean is constantly* θ. *Let* $y_i = a$ *if* $x_i - \theta \geqslant 0$ *and* $y_i = -b$ *if* $x_i - \theta < 0$, *and choose* a, b (>0) *so that* $E(y_i \mid \theta) = 0$ $(i = 1, \dots, n)$. *Use as the test statistic* $m = \max_{0 \leqslant r \leqslant n} \{S_r - \min_{0 \leqslant i < r} S_i\}$, *where* $S_r = \sum_{j=1}^{r} y_j$, $S_0 = 0$, *taking large values as significant, i.e. reject the hypothesis if* $m \geqslant h$.

For simplicity we shall consider only the case where the distribution of the x_i is symmetrical, so that we can take $a = b = 1$; hence $y_i = \operatorname{sgn}(x_i - \theta)$. In order to evaluate the properties of the test let $m_r = S_r - \min_{0 \leqslant i < r} S_i$ and let $p_{r,i}$ be the probability that $m_r = i$ $(i = 0, 1, \dots, h-1)$ and that $m_s < h$ for all s, $1 \leqslant s < r$. Then

$$1 - \sum_{i=1}^{h-1} p_{n,i} \tag{1}$$

is the probability that the null hypothesis is rejected. Let prob $(y_i = 1)$ be $p = 1 - q$. By considering the result of the next observation we have the relations

$$\left.\begin{aligned} p_{r+1,0} &= q(p_{r,0} + p_{r,1}), \\ p_{r+1,i} &= p \cdot p_{r,i-1} + q \cdot p_{r,i+1} \quad (1 \leqslant i < h-1), \\ p_{r+1,h-1} &= p \cdot p_{r,h-2}. \end{aligned}\right\} \tag{2}$$

In matrix notation we have

$$\mathbf{p}_{r+1} = \mathbf{P} \cdot \mathbf{p}_r, \tag{3}$$

where $\mathbf{P}$ is the square matrix,

$$\mathbf{P} = \begin{pmatrix} q & q & 0 & 0 & \dots & 0 & 0 \\ p & 0 & q & 0 & \dots & 0 & 0 \\ 0 & p & 0 & q & \dots & 0 & 0 \\ & & & \dots\dots\dots & & & \\ 0 & 0 & 0 & 0 & \dots & p & 0 \end{pmatrix}. \tag{4}$$

Initially $p_{0,0} = 1$, $p_{0,i} = 0$ $(i \neq 0)$. In this formulation we are implicitly using the fact that the m_r are variables in a Markov chain with h live states. Clearly

$$\mathbf{p}_r = \mathbf{P}^r \cdot \mathbf{p}_0. \tag{5}$$

The expression for $\mathbf{p}_r$ may alternatively be written in terms of the latent roots and vectors, but the simplicity of the matrix $\mathbf{P}$ makes it quite convenient to use (4) for calculation. On the null hypothesis $p = \frac{1}{2}$ constantly. Table 1 shows values of h and the sample size n for which the probabilities of errors of the first kind are at most α, where $\alpha = 0{\cdot}05$ and $0{\cdot}01$. In order to ensure that the Type I errors are at most α for non-tabular values of n the larger value of h should be taken. For larger values of n rough interpolation will provide a sufficiently accurate value of h.

The power of the test II depends both on the value of p after the change and the position of the change. If prob $(y_i = +1)$ is constantly equal to p so that the change can be considered as having occurred immediately, the probability that the null hypothesis is rejected in a sample of n observations is given by equations (1) and (5) with $r = n$. If the change occurs after the kth observation the value of $\mathbf{p}_n$ to be used in (1) is given by

$$\mathbf{p}_n = \mathbf{P}_2^{n-k} \cdot \mathbf{P}_1^k \cdot \mathbf{p}_0,$$

where $\mathbf{P}_1$, $\mathbf{P}_2$ are the matrix $\mathbf{P}$ with $p = \frac{1}{2}$, $p = p$, respectively.

Table 1. *Values of n and h*

$\alpha = 0{\cdot}05$				$\alpha = 0{\cdot}01$	
n	h	n	h	n	h
21	10	75	19	20	12
26	11	83	20	27	14
31	12	91	21	35	16
36	13	100	22	43	18
41	14	119	24	53	20
47	15	139	26	64	22
54	16	161	28	76	24
60	17	185	30	89	26
67	18			103	28
				118	30

In Table 2 the power of the test II is compared with the power of the simple binomial test with approximately the same probability of Type I errors for a sample of 50 observations. For test II to have probability of Type I errors just less than 0·05 we need $h = 16$. The corresponding single sample test is 'Reject the null hypothesis if more than 31 of the y_i are positive', i.e. if $\sum_{i=1}^{50} y_i > 12$. The loss of power from using Test II instead of the single sample test is remarkably small.

Table 2. *Powers of the tests for different p*

p	Test II	Single sample test
0·50	0·039	0·032
0·55	0·136	0·127
0·60	0·336	0·336
0·65	0·609	0·622
0·70	0·844	0·859
0·75	0·964	0·971
0·80	0·996	0·997

The same two tests are contrasted in Table 3, where their powers are shown for different positions of the change from $p = 0{\cdot}5$ to $p = 0{\cdot}75$. Here, however, the test II has an appreciably greater power than the single-sample test when the change occurs near the middle of the set of observations. Also shown in Table 3 are the powers of the single-sample test on the last $50 - m$ observations, when it is known that the change has occurred immediately before the mth observation. The differences between the power of this test and that of test II gives an indication of what is lost from the ignorance of the position of change.

In order to illustrate the test we give an example constructed from tables of random normal deviates. A sample of forty observations was constructed, the first twenty having mean 5 and unit variance, and the last twenty having mean 6 and unit variance; these are shown in Table 4. Suppose that it is required to test the hypothesis that the mean is constantly 5 against the alternatives that an increase in the mean has occurred within the sample. The observations, x_i, are shown in Table 4 together with $y_i = \operatorname{sgn}(x_i - 5)$, and the value taken by $S_r - \min S_t$.

The greatest value, h, of $S_r - \min S_t$ in the sample of 40 is 17, which approaches the 1 % significance point given in Table 1 (for $n = 40$, the approximate 5 % point is $h = 14$, the approximate 1 % point is $h = 18$). This significance level can be compared with that obtained from other tests applied to the

Table 3. *Powers of the tests for different positions of the change*

m	Test II	Single-sample test on whole sample	Single-sample test, m known
0	0·964	0·971	0·971
10	0·906	0·864	0·946
20	0·733	0·625	0·894
30	0·398	0·330	0·618
40	0·122	0·122	0·244
50	0·039	0·032	0·032

Table 4. *Artificial sampling experiment*

Observation no.	1	2	3	4	5	6	7	8	9	10
Value of x_i	3·95	5·96	6·22	5·58	4·02	4·97	3·46	4·29	4·65	5·66
$y_i = \text{sgn}(x_i - 5)$	−1	+1	+1	+1	−1	−1	−1	−1	−1	+1
$S_r - \min S_i$	0	1	2	3	2	1	0	0	0	1
Observation no.	11	12	13	14	15	16	17	18	19	20
Value of x_i	5·44	5·91	4·98	3·58	5·26	3·98	4·19	6·66	6·05	5·97
$y_i = \text{sgn}(x_i - 5)$	+1	+1	−1	−1	+1	−1	−1	+1	+1	+1
$S_r - \min S_i$	2	3	2	1	2	1	0	1	2	3
Observation no.	21	22	23	24	25	26	27	28	29	30
Value of x_i	7·14	6·22	4·76	6·60	5·72	4·88	5·44	5·03	5·66	5·56
$y_i = \text{sgn}(x_i - 5)$	+1	+1	−1	+1	+1	−1	+1	+1	+1	+1
$S_r - \min S_i$	4	5	4	5	6	5	6	7	8	9
Observation no.	31	32	33	34	35	36	37	38	39	40
Value of x_i	6·37	6·66	5·10	5·80	6·29	5·49	4·93	6·18	8·29	6·84
$y_i = \text{sgn}(x_i - 5)$	+1	+1	+1	+1	+1	+1	−1	+1	+1	+1
$S_r - \min S_i$	10	11	12	13	14	15	14	15	16	17

sample. The single-sample binomial test on the y's has 26 positives, 14 negatives; on the null hypothesis. the probability of this or a larger number of positives is 0·04. The change in the mean causes the estimate of variance of the x's to be inflated, and a t-test fails to give significance. The computation required by test II is so simple that it is unnecessary to record the y's, or even the $S_r - \min S_i$. An additional advantage of the test is that it gives an indication where the change took place; the position of the last zero of $S_r - \min S_i$ is an estimate (of course, biased) of the position of change. Thus in the example we would suspect that the change had occurred near observation 17.

3. General remarks

In this section we comment briefly on some other possible methods for the problem of § 2 and related problems without investigating their properties.

Another test for a change in one direction of a parameter from a known specified value may be obtained by analogy with the standard control chart process inspection scheme. The sample is divided into a

number of subsamples of equal size and a statistic calculated from each subsample; the hypothesis of no change is rejected unless all the statistics fall within a certain range. The properties of this test are easy to evaluate and the number of subsamples and the permissible interval for the statistics can be chosen to control the errors. The test is also easy to apply and it is frequently useful in rough work. However the temporal ordering of the observations enters only into the division into subsamples, and it is of interest to examine whether it is advantageous to employ a slightly more complicated test of the form 'Reject the null hypothesis if any k of the statistics calculated from m consecutive subsamples fall outside an interval I, or if any one falls outside a wider interval I'' (cf. Wilkinson, 1951; Tippett, 1931).

The control chart procedure can also provide a test for the two-sided case where the change from the known value can be in either direction. A test based on the mean path of a cumulative sum similar to test I is a truncated sequential test (Rao, 1950). Another case that needs to be considered is where the initial value of the parameter is unknown.

I wish to thank Dr D. R. Cox for a number of discussions on the subject of this paper, and the Director, Mathematical Laboratory, Cambridge, for permission to use the EDSAC.

REFERENCES

Page, E. S. (1954). *Biometrika*, **41**, 100.
Rao, C. R. (1950). *Sankyhā*, **10**, 361.
Tippett, L. H. C. (1931). *The Methods of Statistics*. London: Williams and Norgate.
Wilkinson, B. (1951). *Psychol. Bull.* **48**, 156.

A paradox in statistical estimation

By ALAN STUART

Division of Research Techniques, London School of Economics

1. Sundrum (1954) has recently shown to be incorrect the intuitive idea that the more efficient of two estimators of a parameter necessarily provides the more powerful test of a hypothesis concerning that parameter. This note discusses a similar paradox which arises in a problem concerned purely with estimation: given an estimator u of a parameter θ in a multiparameter distribution, one does not necessarily improve its efficiency by substituting true parameter values into u to replace estimators of them.

2. We shall consider the case where there is only one other parameter, say μ, and where we have a consistent estimator of θ

$$t = f(s, \mu), \tag{1}$$

where s is a function of the n observations only. If μ is unknown, we reduce t to a function of the observations only by substituting for μ a consistent estimator of it, say m, giving

$$u = f(s, m). \tag{2}$$

From (1) we have, to order n^{-1},

$$V(t) = \left(\frac{\partial t}{\partial s}\right)^2 V(s). \tag{3}$$

From (2) we have the corresponding result for a function of two random variables

$$V(u) = \left(\frac{\partial u}{\partial s}\right)^2 V(s) + \left(\frac{\partial u}{\partial m}\right)^2 V(m) + 2\frac{\partial u}{\partial s}\frac{\partial u}{\partial m} C(s, m). \tag{4}$$

Since all the derivatives in (3) and (4) are to be taken at the true parameter point (θ, μ), the first term on the right of (4) is equal to (3). Thus

$$V(u) - V(t) = \left(\frac{\partial u}{\partial m}\right)^2 V(m) + 2\frac{\partial u}{\partial s}\frac{\partial u}{\partial m} C(s, m). \tag{5}$$

(5) is not generally positive, although it must be so if s and m are uncorrelated. In general, their correlation must be taken into account before the effect of substituting parameters into u can be assessed. It is the correlation term in (5) which resolves the paradox.

3. An example in which (5) is negative is provided by the estimation of the correlation parameter ρ in a bivariate normal distribution. A consistent estimator is provided by the sample correlation coefficient r, which has a large-sample variance,

$$V(r) = \frac{1}{n}(1-\rho^2)^2. \tag{6}$$

Although r is the maximum-likelihood estimator of ρ when all five population parameters are being simultaneously estimated, it is not an efficient estimator when the population means and variances are known, which is the case we shall consider. One might therefore expect to improve its efficiency by substituting the known population means and variances for their sample correspondents. The new estimator thus obtained is

$$r' = \frac{\frac{1}{n}\sum_i (x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2}. \tag{7}$$

We find

$$E(r') = \rho, \tag{8}$$

$$n^2\sigma_1^2\sigma_2^2 E\{(r')^2\} = E\{\sum_i (x_i-\mu_1)^2(y_i-\mu_2)^2 + \sum_{i\neq j}(x_i-\mu_1)(y_i-\mu_2)(x_j-\mu_1)(y_j-\mu_2)\}$$

$$= n\mu_{22} + n(n-1)\rho^2\sigma_1^2\sigma_2^2. \tag{9}$$

Since $\mu_{22} = (1+2\rho^2)\sigma_1^2\sigma_2^2$, we have from (8) and (9)

$$V(r') = \frac{1}{n}(1+\rho^2), \tag{10}$$

and comparing (10) with (6) we see that, far from improving the accuracy of estimation, the substitution of true parameter values for random variables in r has multiplied efficiency by a factor $\frac{(1-\rho^2)^2}{1+\rho^2}$, which tends to zero for large ρ^2, and is 1 only when $\rho^2 = 0$.

4. This paradoxical effect can be examined through (5) by restricting our attention to the case when both population means are zero and both variances equal to σ^2, which remains unknown. Our estimator (7) then becomes

$$t = \frac{\frac{1}{n}\sum_i x_i y_i}{\sigma^2},$$

and with a common population variance, we should in our correlation coefficient use a pooled estimator of σ^2, giving

$$u = \frac{\frac{1}{n}\sum_i x_i y_i}{\frac{1}{2n}\sum_i (x_i^2+y_i^2)}.$$

We shall see below that u has the same efficiency as r.

In the notation of (1) and (2) we have

$$t = \frac{s}{\mu} \quad \text{and} \quad u = \frac{s}{m},$$

where

$$s = \frac{1}{n}\sum_i x_i y_i, \quad \mu = \sigma^2 \quad \text{and} \quad m = \frac{1}{2n}\sum_i (x_i^2+y_i^2).$$

We now require

$$\left.\begin{aligned} V(m) &= \frac{\sigma^4}{n}(1+\rho^2), \\ C(s,m) &= \frac{2\rho\sigma^4}{n}. \end{aligned}\right\} \tag{11}$$

Also

$$\left.\begin{aligned} \left(\frac{\partial u}{\partial m}\right)_{\mu,\theta} &= \left(-\frac{u}{m}\right)_{\mu,\theta} = -\frac{\rho}{\sigma^2}, \\ \left(\frac{\partial u}{\partial s}\right)_{\mu,\theta} &= \left(\frac{1}{m}\right)_{\mu} = \frac{1}{\sigma^2}. \end{aligned}\right\} \tag{12}$$

Substituting (11) and (12) into (5), we find

$$V(u)-V(t)=\frac{\rho^2\sigma^4}{\sigma^4}\frac{1}{n}(1+\rho^2)+2\left(\frac{1}{\sigma^2}\right)\left(-\frac{\rho}{\sigma^2}\right)\frac{2\rho\sigma^4}{n}$$

$$=\frac{\rho^2}{n}(\rho^2-3). \tag{13}$$

(13) is negative whenever $\rho^2 \neq 0$, confirming the result on efficiency given above.

5. Finally, we may confirm that u is as efficient as r. Using the general formula for the large-sample variance of a ratio estimator,

$$V(u)=\{E(u)\}^2\left\{\frac{V(s)}{\{E(s)\}^2}+\frac{V(m)}{\{E(m)\}^2}-2\frac{C(s,m)}{E(s)\,E(m)}\right\}. \tag{14}$$

Using (8), (10) and (11) in (14), we obtain

$$V(u)=\rho^2\left\{\frac{1}{n}\frac{(1+\rho^2)}{\rho^2}+\frac{1}{n}(1+\rho^2)-2\frac{2}{n}\right\}$$

$$=\frac{1}{n}(1-\rho^2)^2, \tag{15}$$

agreeing with (6).

REFERENCE

SUNDRUM, R. M. (1954). On the relation between estimating efficiency and the power of tests. *Biometrika*, **41**, 542–4.

Cumulants of a transformed variate

BY G. S. JAMES

University of Leeds

1. Suppose that x is a variate whose ρth cumulant, denoted by $\kappa_{\rho x}$, is of order $\nu^{-\rho+1}$, where ν is some 'large' number. One consequence of this is that x is approximately normally distributed with a variance of order ν^{-1}. Thus if y is some well-behaved function of x, say $y=f(x)$, then y will also be approximately normally distributed with mean and variance given by

$$\kappa_{1y}=\mathscr{E}y=f(\mu)\,[1+O(\nu^{-1})], \tag{1}$$

$$\kappa_{2y}=\operatorname{var}y=[f'(\mu)]^2\operatorname{var}x[1+O(\nu^{-1})], \tag{2}$$

where μ is the mean value of x. These are of orders ν^0 and ν^{-1} respectively.

It is by no means obvious, however, that the ρth cumulant of y, denoted by $\kappa_{\rho y}$, is of order $\nu^{-\rho+1}$ for general values of ρ, and not merely for $\rho=1$ and 2. The object of this note is to show that this is true, at any rate formally. (The result will be true in reality, as well as formally, under the same sort of conditions for which (1) and (2) are really true.)

Suppose that $f(x)$ can be expanded formally in a Taylor series round μ; there is no loss of generality in taking $\mu=0$, so that this series is

$$y=c_0+c_1x+c_2x^2+\ldots, \tag{3}$$

where we suppose that $c_0, c_1, c_2, \ldots$ do not depend upon ν. It is quite easy to show, by formal expansion starting from (3), that the ρth moment of y about its mean is of order $\nu^{-[\frac{1}{2}(\rho+1)]}$, where $[k]$ denotes the integral part of k. But on conversion to cumulants the higher order terms always seem to cancel, to such an extent that $\kappa_{\rho y}$ is of order $\nu^{-\rho+1}$. We now give a proof of this result.

THEOREM 1. If a variate x possesses cumulants $\kappa_{\rho x}$ of all orders, and if $\kappa_{\rho x}=O(\nu^{-\rho+1})$ $(\rho=1,2,\ldots)$, and the cumulants of $y=f(x)$ are calculated on the basis of a (possibly formal) Taylor expansion (3), where the c_r do not depend upon ν, then $\kappa_{\rho y}=O(\nu^{-\rho+1})$ $(\rho=1,2,\ldots)$.

Without loss of generality we may assume that $c_0=0$. We may then write $y=\Sigma z_r$, where $z_r=c_rx^r$ $(r=1,2,\ldots)$. Writing $\mu'_{(s)}(r_1\ldots r_\rho)$ for the moment $\mathscr{E}z_{r_1}\ldots z_{r_\rho}$ (some of the r_h may be equal) and $\kappa_{(s)}(r_1\ldots r_\rho)$

for the corresponding cumulant, we have, by the properties of moment-generating and cumulant-generating functions,

$$\frac{\kappa_{1\nu}t}{1!}+\frac{\kappa_{2\nu}t^2}{2!}+\ldots=\log\left[1+\frac{t\mathscr{E}y}{1!}+\frac{t^2\mathscr{E}y^2}{2!}+\ldots\right]$$

$$=\log\left[1+\frac{1}{1!}\Sigma t_r\mu'_{(s)}(r)+\frac{1}{2!}\Sigma t_r t_s\mu'_{(s)}(rs)+\ldots\right]$$

$$=\frac{1}{1!}\Sigma t_r\kappa_{(s)}(r)+\frac{1}{2!}\Sigma t_r t_s\kappa_{(s)}(rs)+\ldots, \tag{4}$$

where $t_1=t_2=\ldots=t$ and the summations are over $r,s,\ldots=1,2,\ldots$. Hence

$$\kappa_{\rho\nu}=\Sigma\kappa_{(s)}(r_1\ldots r_\rho), \tag{5}$$

summed for $r_1,\ldots,r_\rho=1,2,\ldots$.

Thus to prove the theorem it is sufficient to show that $\kappa_{(s)}(r_1\ldots r_\rho)$ is of order $\nu^{-\rho+1}$ or lower. To demonstrate this, suppose rather more generally that $x_1,\ldots,x_n$ is a sample of n independent values from the x distribution, and let $z_r=c_r(\Sigma x_i)^r$ $(r=1,2,\ldots)$. This agrees with the earlier definition of z_r when $n=1$ (and $x_1=x$). The z_r can be regarded as statistics of the sample $x_1,\ldots,x_n$, although of course they are functionally related. Each z_r is a homogeneous polynomial symmetric function of the sample values, of degree r, and the coefficient of $x_1^{a_1}\ldots x_\alpha^{a_\alpha}$ in z_r (with $\Sigma a_j=r$) is

$$\frac{r!}{a_1!\ldots a_\alpha!}B_{(s)}(a_1\ldots a_\alpha), \tag{6}$$

where

$$B_{(s)}(a_1\ldots a_\alpha)=c_r. \tag{7}$$

A corresponding result holds for Fisher's statistics k_r, but with

$$B_{(k)}(a_1\ldots a_\alpha)=\frac{(-)^{\alpha-1}(\alpha-1)!}{n(n-1)\ldots(n-\alpha+1)}. \tag{8}$$

The factors $B_{(k)}$ and $B_{(s)}$ happen not to depend on the complete detail of the partition $(a_1\ldots a_\alpha)$ of the number r; but I have shown (James, 1955) that Fisher's rules for obtaining the sampling cumulants of k-statistics, described in Kendall's book (Kendall, 1947), may be adapted to *any* system of statistics $z_1,z_2,\ldots$ (z_r being a homogeneous polynomial symmetric function of degree r in the observations) by merely replacing the factors $B_{(k)}$ occurring in the evaluation of the 'pattern functions' by the factors $B_{(s)}$, where $B_{(s)}$ is the coefficient of $x_1^{a_1}\ldots x_\alpha^{a_\alpha}$ in z_r divided by the appropriate multinomial coefficient.

The rule of Fisher which is of paramount importance for our proof is the one which states that, in finding cumulants of k- (or z-) statistics in terms of population cumulants, we are to neglect any array which splits up into two or more disjoint blocks. Now any array which does *not* fall into disjoint blocks may be built up column by column in such a way that at each stage the new column does not form an isolated block by itself. (For example, the columns of $\begin{matrix}\times & \times\times\\ \times & \times\times\end{matrix}$ may be taken in the order 1324, but not in the order 1234). Thus, if $\alpha_1,\ldots,\alpha_\rho$ are the numbers of non-zero elements in the 1st, ..., ρth columns and τ is the total number of rows in the array, we have

$$\tau\leqslant\alpha_1+(\alpha_2-1)+\ldots+(\alpha_\rho-1)=\Sigma\alpha_h-\rho+1. \tag{9}$$

Now if this array is one of those contributing to the coefficient of $\kappa_{t_1}\ldots\kappa_{t_\tau}$ in $\kappa_{(s)}(r_1\ldots r_\rho)$ (so that $\Sigma t_j=\Sigma r_h$) then the corresponding term is of order $\nu^{\Sigma(-t_j+1)}=\nu^{-\Sigma r_h+\tau}$; for the numerical factors and pattern functions for each separation of the array do not depend upon ν, while each κ_{t_j} is of order ν^{-t_j+1}. But (9) shows that

$$-\Sigma r_h+\tau\leqslant-\Sigma(r_h-\alpha_h)-\rho+1\leqslant-\rho+1; \tag{10}$$

for α_h is the number of parts in a partition of r_h, and so does not exceed r_h. Thus this particular term of $\kappa_{(s)}(r_1\ldots r_\rho)$ is of order $\nu^{-\rho+1}$ or lower. Hence the same is true of $\kappa_{(s)}(r_1\ldots r_\rho)$ itself, and finally, by (5), of $\kappa_{\rho\nu}$.

2. By the use of multivariate sampling rules more general results than Theorem 1 can be proved. Perhaps the most general is Theorem 2.

THEOREM 2. If $y^1=f^1(x^1,\ldots,x^p)$, $y^2=f^2(x^1,\ldots,x^p),\ldots$ are functions of the variates* $x^1,\ldots,x^p$, formally expansible in the forms

$$y^h=c^h+\Sigma c_i^h x^i+\Sigma c_{ij}^h x^i x^j+\ldots, \tag{11}$$

* The superfixes are indices, not exponents.

and if the ρth-order cumulant, $\kappa_x^{i_1 \dots i_\rho}$, of $x^{i_1}, \dots, x^{i_\rho}$ is of order $\nu^{-\rho+1}$ for $i_1, \dots, i_\rho = 1, \dots, p$ and $\rho = 1, 2, \dots$, then the same holds for the cumulants, $\kappa_y^{\lambda_1 \dots \lambda_\rho}$, of the y^λ. (Here we use a slightly modified form of the 'tensor' notation suggested by Kaplan (1952).)

An outline of the proof is as follows. If $z_1, z_2, \dots$ denote the quantities $x^1, \dots, x^p, x^1x^1, x^1x^2, \dots, x^px^p, x^1x^1x^1, \dots$ written in some convenient order, then (11) may be rewritten in the form

$$y^\lambda = \Sigma c^{\lambda s} z_s, \tag{12}$$

whence we easily derive

$$\kappa_y^{\lambda_1, \dots \lambda_\rho} = \sum_{s_1, \dots, s_\rho} c^{\lambda_1 s_1} \dots c^{\lambda_\rho s_\rho} \kappa_{(z)}(s_1 \dots s_\rho), \tag{13}$$

where $\kappa_{(z)}(s_1 \dots s_\rho)$ is the mixed cumulant of $z_{s_1}, \dots, z_{s_\rho}$. Thus it suffices to show that all the $\kappa_{(z)}(s_1 \dots s_\rho)$ are of order $\nu^{-\rho+1}$ or lower, for $\rho = 1, 2, \dots$. The proof is completed in much the same way as before, but in order to see clearly the application of the results of my other paper the z's should be relabelled as follows: $z^i = x^i$, $z^{ij} = x^ix^j$, ...

3. *An example.* Johnson & Welch (1939) have found by direct calculation, using Stirling's asymptotic expansion of the gamma function, that if χ^2 is the sum of squares of ν independent standard normal deviates, then the cumulants of χ, as far as the sixth, are given asymptotically as follows:

$$\left.\begin{aligned} &\kappa_{1\chi} \sim \nu^{\frac{1}{2}}, && \kappa_{2\chi} \sim \tfrac{1}{2}, \\ &\kappa_{3\chi} \sim \tfrac{1}{4}\nu^{-\frac{1}{2}}, && \kappa_{4\chi} \sim \tfrac{3}{16}\nu^{-1}, \\ &\kappa_{5\chi} \sim -\tfrac{3}{16}\nu^{-\frac{3}{2}}, && \kappa_{6\chi} \sim -\tfrac{15}{16}\nu^{-2}. \end{aligned}\right\} \tag{14}$$

Now if we write $x = \chi^2/\nu - 1$, $y = \sqrt{(1+x)} = \chi/\sqrt{\nu}$, then $\kappa_{\rho x}$ is of order $\nu^{-\rho+1}$, so that Theorem 1 shows that $\kappa_{\rho y}$ is of order $\nu^{-\rho+1}$ or lower and $\kappa_{\rho\chi}$ is of order $\nu^{-\frac{1}{2}\rho+1}$ or lower. Thus the odd cumulants in (14) are of the order we should expect, but the even ones (apart from the second) seem to be a whole order lower than demanded by Theorem 1. I have been unable to prove that $\kappa_{2\rho,\chi} = O(\nu^{-\rho})$ for general values of ρ ($\geqslant 2$).

REFERENCES

James, G. S. (1955). On moments and cumulants of systems of statistics. (Not yet published.)

Johnson, N. L. & Welch, B. L. (1939). On the calculation of the cumulants of the χ-distribution. *Biometrika*, **31**, 216–18.

Kaplan, E. L. (1952). Tensor notation and the sampling cumulants of k-statistics. *Biometrika*, **39**, 319–23.

Kendall, M. G. (1947). *The Advanced Theory of Statistics*, vol. **1**, 3rd ed. London: Charles Griffin and Co.

The likelihood ratio test for Markoff chains

By I. J. Good

1. The present paper is virtually a footnote to that of Hoel (1954), who, using and acknowledging the methods of Bartlett (1951), constructed a likelihood ratio test for the order of a Markoff chain. If H_ν is the hypothesis that the chain is of order ν, then Hoel tests $H_{\nu-1}$ against (or rather *within*) H_ν. Here we make a simple generalization so as to test H_μ within H_ν, and we relate the work to some previous work.

2. Let $x_1, x_2, \dots, x_N$ be a sequence, $\mathfrak{S}$, of observations, each observation being capable of taking one of t values denoted conventionally by $1, 2, \dots, t$. Let H_ν be the hypothesis that $\mathfrak{S}$ is a Markoff chain of order ν ($\nu = 0, 1, 2, \dots$). Note that H_0 means that $\mathfrak{S}$ is a random sequence. We may obtain uniformity in expression by defining H_{-1} to mean that $\mathfrak{S}$ is a 'perfectly' random or 'equiprobably' random sequence. H_{-1} is the only H_μ that is a simple statistical hypothesis. Clearly H_μ implies H_ν whenever $\mu < \nu$.

The likelihood ratio test for composite hypotheses may be expressed in the following form. Let H and H' be hypotheses such that H implies H', so that H' at least is not a simple statistical hypothesis. Let E be an experimental result and let

$$\lambda = \max P(E \mid H^*) / \max P(E \mid H'^*),$$

the maxima being taken over all simple statistical hypotheses H^*, H'^* belonging to H and H' respectively. λ is the likelihood ratio statistic for composite hypotheses and we may say that it tests H *within* H'. We may call H the null hypothesis whether or not it is simple and (within orthodox statistics) λ can be used for the rejection of H.

Hoel (1954) used λ for testing $H_{\nu-1}$ within H_ν ($\nu \geqslant 1$), and, with suitable conventions, his result applies also for $\nu = 0$. Bartlett (1951) used the likelihood ratio statistic for testing completely specified chains of order $\nu - 1$ within H_ν. Since H_{-1} can be regarded as a completely specified chain of *any* order, Bartlett's work included tests for H_{-1} within H_ν ($\nu = 0, 1, 2, \ldots$). Similarly, a completely specified chain of order μ is also one of order $\nu - 1$ (if $\mu < \nu$), so that Bartlett's work applies for testing any completely specified chain of order μ within H_ν where $\mu < \nu$.

We shall here generalize Hoel's work in order to test H_μ within H_ν where $\mu < \nu$. Only when $\mu = -1$ does the generalization overlap with Bartlett's results, since the only H_μ that is completely specified is H_{-1}. (A completely specified statistical hypothesis is the same thing as a simple statistical hypothesis.)

3. Associated with $\mathfrak{S}$ is the corresponding cyclic sequence $\overline{\mathfrak{S}}$ defined by regarding the first element of $\mathfrak{S}$ as immediately following its last one. We shall always denote properties of $\overline{\mathfrak{S}}$ by placing a bar over the corresponding algebraic symbol relating to $\mathfrak{S}$.

A sequence of ν consecutive observations is called a *ν-sequence*. Let $n_{r_1,\ldots,r_\nu}$, or $n_{\mathbf{r}}$ for short, be the number of ν-sequences in $\mathfrak{S}$ which are $(r_1, r_2, \ldots, r_\nu) = \mathbf{r}$. Let $\overline{n}_{\mathbf{r}}$ be defined similarly. Let

$$\psi_\nu^2 = \sum_{\mathbf{r}} \left(n_{\mathbf{r}} - \frac{N-\nu+1}{t^\nu} \right)^2 \Big/ \frac{N-\nu+1}{t^\nu} \quad (\nu \geqslant 1), \tag{1}$$

$$\overline{\psi}_\nu^2 = \sum_{\mathbf{r}} (\overline{n}_{\mathbf{r}} - Nt^{-\nu})^2 / Nt^{-\nu} \quad (\nu \geqslant 1), \tag{2}$$

$$\psi_0^2 = \overline{\psi}_0^2 = 0, \tag{3}$$

$$K_\nu = 2 \sum_{\mathbf{r}} n_{\mathbf{r}} \log n_{\mathbf{r}} \quad (\nu \geqslant 1), \tag{4}$$

$$K_0 = 2(N-\nu+1)\log(N-\nu+1), \quad K_{-1} = 0, \tag{5}$$

$$\overline{K}_\nu = 2 \sum_{\mathbf{r}} \overline{n}_{\mathbf{r}} \log \overline{n}_{\mathbf{r}} \quad (\nu \geqslant 1), \tag{6}$$

$$\overline{K}_0 = 2N \log N, \quad \overline{K}_{-1} = 0, \tag{7}$$

$$\nabla\psi_\nu^2 = \psi_\nu^2 - \psi_{\nu-1}^2, \quad \text{etc.}, \tag{8}$$

$$\nabla^2\psi_\nu^2 = \psi_\nu^2 - 2\psi_{\nu-1}^2 + \psi_{\nu-2}^2, \quad \text{etc.} \tag{9}$$

(The logarithms are to base e and $0 \log 0$ means 0.)

It is known that if H_{-1} is true, then ψ_ν^2 and $\overline{\psi}_\nu^2$ do not have asymptotically gamma variate (chi-squared) distributions, but $\nabla\psi_\nu^2$, $\nabla\overline{\psi}_\nu^2$, $\nabla^2\psi_\nu^2$, and $\nabla^2\overline{\psi}_\nu^2$ do (see Bartlett, 1951; Good, 1953). Moreover (precisely),

$$\mathscr{E}\psi_\nu^2 = \mathscr{E}\overline{\psi}_\nu^2 = t^\nu - 1, \tag{10}$$

while $\nabla\psi_\nu^2$ and $\nabla\overline{\psi}_\nu^2$ have ∇t^ν degrees of freedom ($\nu \geqslant 1$), and $\nabla^2\psi_\nu^2$ and $\nabla^2\overline{\psi}_\nu^2$ have $\nabla^2 t^\nu$ degrees of freedom ($\nu \geqslant 2$).

Hoel defined λ as the ratio of the maximum likelihood given $H_{\nu-1}$ to that given H_ν and obtained the asymptotic distribution of λ, given $H_{\nu-1}$. Let us denote by $\lambda_{\mu,\nu}$ the ratio of the maximum likelihood given H_μ to that given H_ν ($\mu < \nu$). Then Hoel's λ is our $\lambda_{\nu-1,\nu}$ and obviously $\lambda_{\mu,\nu}$ is the product of $\nu - \mu$ of Hoel's λ's, namely

$$\lambda_{\mu,\nu} = \lambda_{\mu,\mu+1}\lambda_{\mu+1,\mu+2}\cdots\lambda_{\nu-1,\nu}. \tag{11}$$

The expression given by Hoel for $-2\log\lambda_{\nu-1,\nu}$ may be written in the form

$$-2\log\lambda_{\nu-1,\nu} = \nabla^2 K_{\nu+1} \quad (\nu = 1, 2, 3, \ldots). \tag{12}$$

With our conventions this equation is also valid for $\nu = 0$.

Hoel's result can be stated in the form: When $H_{\nu-1}$ is given, $\nabla^2 K_{\nu+1}$ has asymptotically a gamma variate distribution with $\nabla^2 t^{\nu+1}$ degrees of freedom if $\nu \geqslant 1$ and we may add that for $\nu = 0$ it has asymptotically a gamma variate distribution with $t - 1$ degrees of freedom, since it then reduces to the likelihood ratio test for 'perfect' (equiprobable) randomness against randomness in general. Clearly

$$\begin{aligned} -2\log\lambda_{\mu,\nu} &= -2\log\lambda_{\mu,\mu+1} - \ldots - 2\log\lambda_{\nu-1,\nu} \\ &= \nabla^2 K_{\nu+1} + \nabla^2 K_\nu + \ldots + \nabla^2 K_{\mu+2} \\ &= \nabla K_{\nu+1} - \nabla K_{\mu+1} \quad (-1 \leqslant \mu < \nu). \end{aligned} \tag{13}$$

If we could assume that the variables $-2\log\lambda_{\nu-1,\nu}$ ($\nu = 0, 1, 2, \ldots$) were asymptotically independent, given H_μ, it would follow that $-2\log\lambda_{\mu,\nu}$ has a gamma variate distribution with degrees of freedom

$$\left.\begin{aligned} \nabla t^{\nu+1} - \nabla t^{\mu+1} \quad & (\mu \geqslant 0), \\ \nabla t^{\nu+1} \quad & (\mu = -1), \end{aligned}\right\} \tag{14}$$

and we could use these results for testing H_μ within H_ν. Unfortunately, the independence does not seem to be easy to prove. Nevertheless, the result just stated is correct and can be proved by precisely the same method as Hoel used, except that his suffix i is to be replaced throughout his proof by a sequence of $\nu-\mu$ suffixes. It is unnecessary to repeat the argument since the modifications are entirely trivial. The conjecture that the variables $-2\log\lambda_{\nu-1,\nu}$ are asymptotically independent is strengthened by the knowledge that the above deduction from it is correct.

4. The value of the generalization of Hoel's results is that there may not be a significant distinction between 'adjacent' hypotheses in the sequence $H_\mu, H_{\mu+1}, \ldots, H_\nu$, yet H_μ may be clearly rejectable by the statistic $\nabla K_{\nu+1} - \nabla K_{\mu+1}$. If we put $\mu = 0$ [$\mu = -1$] we have a test for randomness ('perfect' randomness) within Markovity of order ν. If $\mu = 0$ and $\nu = 1$ we obtain a test that is a special case of the likelihood ratio test for contingency tables. (See, for example, Wilks, 1946, p. 220, where, however, there is a minor slip in that the expression given as λ is really $1/\lambda$.)

5. There is clearly a strong analogy between the expressions $\nabla K_{\nu+1} - \nabla K_{\mu+1}$ and $\nabla\psi^2_{\nu+1} - \nabla\psi^2_{\mu+1}$. When H_{-1} is true the latter expression also has asymptotically a gamma variate distribution with the same number of degrees of freedom as the former expression, since (as shown by Good (1953)) the variables $\nabla^2\psi^2_\nu$ ($\nu = 1, 2, \ldots$) are asymptotically independent, at any rate when t is a prime number. This analogy is no coincidence since $\nabla\psi^2_\nu$ is the asymptotic form of $\nabla K_{\nu+1} - \nabla K_0$ when H_{-1} is true.

6. We may also use the cyclic definitions. When H_μ is true, $\nabla\bar{K}_{\nu+1} - \nabla\bar{K}_{\mu+1}$ has asymptotically a gamma variate distribution with a number of degrees of freedom given by (14). The cyclic definition is mathematically simpler than the non-cyclic one and makes the checks

$$\sum_{r_\nu} n_{\mathbf{r}} = n_{r_1,\ldots,r_{\nu-1}}, \qquad \sum_{r_1} n_{\mathbf{r}} = n_{r_2 \ldots r_\nu}$$

precise.

7. Not much is known concerning the accuracy of the asymptotic formulae when N is specified. When testing H_{-1} the psi-squared statistic has the advantage that its expected value is precisely known, but in principle this statistic may well be less powerful than the likelihood ratio. Unless N is very small there is probably little to choose between the cyclic and non-cyclic forms of the latter.

8. Another statistic that would be worth consideration would be $\nabla L_{\nu+1} - \nabla L_{\mu+1}$ (and its cyclic form), where

$$L_r = 2\sum_{\mathbf{r}} \log n_{\mathbf{r}}! \quad (\nu \geqslant 1), \qquad L_0 = 2\log N!, \qquad L_{-1} = 0. \tag{15}$$

When $\nu = 1$ and $\mu = 0$ this statistic reduces to $\nabla^2 L_2$, which is minus twice the log likelihood of the n_{r_1,r_2}'s, regarded as forming the interior of a contingency table for which the marginal totals are assigned and independence is assumed.

REFERENCES

Bartlett, M. S. (1951). The frequency goodness of fit for probability chains. *Proc. Camb. Phil. Soc.* **47**, 86–95.

Good, I. J. (1953). The serial test for sampling numbers and other tests for randomness. *Proc. Camb. Phil. Soc.* **49**, 276–84.

Hoel, P. G. (1954). A test for Markoff chains. *Biometrika*, **41**, 430–3.

Wilks, S. S. (1946). *Mathematical Statistics*. Princeton.

Exact forms of some invariants for distributions admitting sufficient statistics

By V. S. HUZURBAZAR

University of Poona, India

1. Introduction

The concept of distance in statistics is due to Mahalanobis (1936) who defined the distance between two multivariate normal populations. Later Jeffreys (1946, 1948) defined a wider class of invariants of probability distributions which may be regarded as providing measures of distance between two

probability distributions, in general. If p and p' are the density functions of two probability distributions of a variate x, then the expressions defined by Jeffreys

$$I_m = \int |\, p'^{1/m} - p^{1/m} \,|^m \, dx, \tag{1}$$

and

$$J = \int (p' - p) \log \frac{p'}{p} \, dx, \tag{2}$$

are all positive definite and invariant for all non-singular transformations of the variate and the parameters. These expressions, therefore, provide measures of distance between two distributions. For discrete distributions p and p' are simply the probabilities that x takes a particular value x. Extension to multivariate distributions follows immediately.

An important case is the distance between two distributions having the *same mathematical form* but with *different* sets of values of the parameters. If the corresponding parameters in the two sets differ by infinitesimals, we get the differential forms of the invariants I_m and J. The invariants I_2 and J have interesting mathematical properties and have been used by Jeffreys in stating the prior probabilities of parameters in his theory of estimation and tests of significance. Though these invariants have not yet attracted the attention of the frequency theorists of probability, it is likely that they may find applications in their work also.

Jeffreys has obtained the exact forms of the invariants I_2 and J for the univariate and bivariate normal distributions, the Poisson and the binomial distributions; and for these distributions the exact forms come out as *explicit* functions of the parameters of distributions.

The object of this paper is to prove, in general terms, a remarkable property that for all distributions admitting sufficient statistics the exact forms of I_m (m even) and J come out as *explicit* functions of the parameters of distributions. It may be mentioned here that the properties of I_m ($m \neq 2$) have so far remained uninvestigated, the form of I_m being then very complicated.

2. Distributions admitting sufficient statistics

We shall take the most general form of distributions admitting sufficient statistics as given by Koopman (1936),

$$f(x, \alpha_j) = \exp\left\{ \sum_{k=1}^{p} u_k(\alpha_j)\, v_k(x) + A(x) + B(\alpha_j) \right\}, \tag{3}$$

where, for brevity, (α_j) denotes the set of p parameters $(\alpha_1, \alpha_2, \ldots, \alpha_p)$; u_k and B are functions of the α_j, and v_k and A are functions of x. For multivariate distributions (x) is to be replaced by the set of variates $(x_1, x_2, \ldots, x_q)$.

Since $\int f(x, \alpha_j)\, dx \equiv 1$, for all α_j, we have

$$\int \exp\left\{ \sum_{k=1}^{p} u_k(\alpha_j)\, v_k(x) + A(x) \right\} dx = \exp\{-B(\alpha_j)\}. \tag{4}$$

Now the $u_k(\alpha_j)$ are p independent functions of the p parameters α_j. We can express the α_j inversely as functions of the u_k's. Then $B(\alpha_j)$ can be expressed in terms of the u_k's as

$$B(\alpha_j) = b(u_k). \tag{5}$$

Then (4) becomes

$$\int \exp\left\{ \sum_{k=1}^{p} u_k(\alpha_j)\, v_k(x) + A(x) \right\} dx = \exp\{-b(u_k)\}. \tag{6}$$

3. Exact form of I_2

Let

$$f(x, \alpha_j) = \exp\left\{ \sum_{k=1}^{p} u_k(\alpha_j)\, v_k(x) + A(x) + B(\alpha_j) \right\}, \tag{7}$$

and

$$f(x, \alpha'_j) = \exp\left\{ \sum_{k=1}^{p} u_k(\alpha'_j)\, v_k(x) + A(x) + B(\alpha'_j) \right\}, \tag{8}$$

so that $f(x, \alpha_j)$ and $f(x, \alpha_j')$ have the same mathematical form but different sets of values of the parameters. We have

$$\begin{aligned} I_2 &= \int \{\sqrt{[f(x,\alpha_j)]} - [f(x,\alpha_j')]\}^2 dx \\ &= 2 - 2\int \{f(x,\alpha_j) f(x,\alpha_j')\}^{\frac{1}{2}} dx \\ &= 2 - 2\int \exp\left[\sum_{k=1}^{p} \{\tfrac{1}{2}u_k(\alpha_j) + \tfrac{1}{2}u_k(\alpha_j')\} v_k(x) + A(x) + \tfrac{1}{2}B(\alpha_j) + \tfrac{1}{2}B(\alpha_j')\right] dx \\ &= 2 - 2\exp\{\tfrac{1}{2}B(\alpha_j) + \tfrac{1}{2}B(\alpha_j')\} \int \exp\left[\sum_{k=1}^{p} \{\tfrac{1}{2}u_k(\alpha_j) + \tfrac{1}{2}u_k(\alpha_j')\} v_k(x) + A(x)\right] dx. \end{aligned} \tag{9}$$

Writing $\frac{1}{2}u_k(\alpha_j) + \frac{1}{2}u_k(\alpha_j')$ for $u_k(\alpha_j)$ in (6), we have

$$\int \exp\left[\sum_{k=1}^{p} \{\tfrac{1}{2}u_k(\alpha_j) + \tfrac{1}{2}u_k(\alpha_j')\} v_k(x) + A(x)\right] dx = \exp[-b\{\tfrac{1}{2}u_k(\alpha_j) + \tfrac{1}{2}u_k(\alpha_j')\}]. \tag{10}$$

From (9) and (10) we have

$$I_2 = 2 - 2\exp[\tfrac{1}{2}B(\alpha_j) + \tfrac{1}{2}B(\alpha_j') - b\{\tfrac{1}{2}u_k(\alpha_j) + \tfrac{1}{2}u_k(\alpha_j')\}]. \tag{11}$$

The curious point to be noted is that the function $A(x)$ remains unaltered in the integral on the left-hand side of (10) which enables us to evaluate that integral explicitly in virtue of (6).

3·1. *Illustrative example*

Consider the Type III distribution

$$f(x, a, p) = \frac{a^p e^{-ax} x^{p-1}}{\Gamma(p)} \quad (0 \leqslant x < \infty). \tag{12}$$

We write

$$f(x, a, p) = \exp\{-ax + p\log x - \log x + p\log a - \log\Gamma(p)\}. \tag{13}$$

Here $u_1 = a$, $u_2 = p$, $B = p\log a - \log\Gamma(p)$. Expressing B in terms of u_1 and u_2,

$$B = u_2 \log u_1 - \log\Gamma(u_2) = b(u_1, u_2),$$

$$b\left(\frac{a+a'}{2}, \frac{p+p'}{2}\right) = \left(\frac{p+p'}{2}\right)\log\frac{a+a'}{2} - \log\Gamma\left(\frac{p+p'}{2}\right),$$

$$\begin{aligned} I_2 &= 2 - 2\exp\left[\tfrac{1}{2}p\log a - \tfrac{1}{2}\log\Gamma(p) + \tfrac{1}{2}p'\log a' - \tfrac{1}{2}\log\Gamma(p') - \left(\frac{p+p'}{2}\right)\log\left(\frac{a+a'}{2}\right) + \log\Gamma\left(\frac{p+p'}{2}\right)\right] \\ &= 2 - \frac{2\Gamma\left(\frac{p+p'}{2}\right)}{\{\Gamma(p)\,\Gamma(p')\}^{\frac{1}{2}}} \frac{a^{\frac{1}{2}p} a'^{\frac{1}{2}p'}}{\left(\frac{a+a'}{2}\right)^{\frac{1}{2}(p+p')}}. \end{aligned} \tag{14}$$

4. The exact form of J

$$\begin{aligned} J &= \int \{f(x,\alpha_j') - f(x,\alpha_j)\}\{\log f(x,\alpha_j') - \log f(x,\alpha_j)\} dx \\ &= \sum_{k=1}^{p} \{u_k(\alpha_j') - u_k(\alpha_j)\}\left\{\int v_k(x) f(x,\alpha_j')\, dx - \int v_k(x) f(x,\alpha_j)\, dx\right\} \\ &= \sum_{k=1}^{p} \{u_k(\alpha_j') - u_k(\alpha_j)\}\{E' v_k(x) - E v_k(x)\}, \end{aligned} \tag{15}$$

where $E'v_k(x)$ and $Ev_k(x)$ denote the expectations of $v_k(x)$ when the parameters are α_j' and α_j respectively. For brevity write

$$E_k' = E' v_k(x), \quad E_k = E v_k(x).$$

Then

$$J = \sum_{k=1}^{p} \{u_k(\alpha_j') - u_k(\alpha_j)\}(E_k' - E_k). \tag{16}$$

E'_k and E_k can be obtained as follows. We have

$$\frac{\partial}{\partial\alpha_r}\log f(x,\alpha_j)=\sum_{k=1}^{p}\frac{\partial u_k}{\partial\alpha_r}v_k(x)+\frac{\partial B}{\partial\alpha_r}.$$

Since $$E\left\{\frac{\partial}{\partial\alpha_r}\log f(x,\alpha_j)\right\}\equiv 0,$$

we have $$\sum_{k=1}^{p}\frac{\partial u_k}{\partial\alpha_r}E_k+\frac{\partial B}{\partial\alpha_r}=0. \tag{17}$$

Setting $r=1,2,\ldots,p$ in (17), we have p simultaneous linear equations to determine the E_k. Then E'_k are obtained by writing α'_j for α_j in E_k. Thus (15) and (17) enable us to express J explicitly in terms of the parameters.

The above formulae are greatly simplified if we take the u_k as parameters and express $B(\alpha_j)$ in terms of them. Then (17) becomes simply

$$E_r+\frac{\partial B}{\partial u_r}=0, \quad \text{so that} \quad E_r=-\frac{\partial B}{\partial u_r}.$$

From (16), $$J=\sum_{k=1}^{p}(u'_k-u_k)\left(-\frac{\partial B'}{\partial u'_k}+\frac{\partial B}{\partial u_k}\right). \tag{18}$$

Writing $u'_k-u_k=du_k$, the differential form of J is

$$J=-\sum_{k=1}^{p}\sum_{i=1}^{p}\frac{\partial^2 B}{\partial u_k\partial u_i}du_k du_i. \tag{19}$$

5. The exact form of I_m (m even)

For brevity write

$$f(x,\alpha_j)=f, \quad f(x,\alpha'_j)=f', \quad u_k(\alpha'_j)=u'_k, \quad B(\alpha'_j)=B', \quad \text{etc.}$$

Now
$$\begin{aligned}I_m&=\int|f'^{1/m}-f^{1/m}|^m\,dx\\&=\int(f'^{1/m}-f^{1/m})^m\,dx \quad (\text{since } m \text{ is even})\\&=\int\left\{\sum_{r=0}^{m}(-1)^r\binom{m}{r}f'^{(m-r)/m}f^{r/m}\right\}dx\\&=\sum_{r=0}^{m}(-1)^r\binom{m}{r}\lambda_r,\end{aligned} \tag{20}$$

where
$$\begin{aligned}\lambda_r&=\int f'^{(m-r)/m}f^{r/m}\,dx\\&=\int\exp\left\{\sum_{k=1}^{p}\left(\frac{m-r}{m}u'_k+\frac{r}{m}u_k\right)v_k(x)+A(x)+\frac{m-r}{m}B'+\frac{r}{m}B\right\}dx\\&=\exp\left(\frac{m-r}{m}B'+\frac{r}{m}B\right)\int\exp\left\{\sum_{k=1}^{p}\left(\frac{m-r}{m}u'_k+\frac{r}{m}u_k\right)v_k(x)+A(x)\right\}dx.\end{aligned} \tag{21}$$

Writing $\frac{m-r}{m}u'_k+\frac{r}{m}u_k$ for u_k in (6), we have

$$\int\exp\left\{\sum_{k=1}^{p}\left(\frac{m-r}{m}u'_k+\frac{r}{m}u_k\right)v_k(x)\,A(x)\right\}dx=\exp\left\{-b\left(\frac{m-r}{m}u'_k+\frac{r}{m}u_k\right)\right\}. \tag{22}$$

From (21) and (22),

$$\lambda_r=\exp\left\{\frac{m-r}{m}B'+\frac{r}{m}B-b\left(\frac{m-r}{m}u'_k+\frac{r}{m}u_k\right)\right\}.$$

Then from (20), $$I_m = \sum_{r=0}^{m} (-1)^r \binom{m}{r} \exp\left\{\frac{m-r}{m} B' + \frac{r}{m} B - b\left(\frac{m-r}{m} u_k' + \frac{r}{m} u_k\right)\right\}. \quad (23)$$

Curiously again, the function $A(x)$ remains unaltered in the integral in (22), which enables us to evaluate that integral explicitly in virtue of (6).

REFERENCES

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc.* A, **186**, 453.

JEFFREYS, H. (1948). *Theory of Probability*, 2nd ed. Oxford: Clarendon Press.

KOOPMAN, B. O. (1936). On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.* **39**, 399.

MAHALANOBIS, P. C. (1936). On the generalized distance in statistics. *Proc. Nat. Inst. Sci. Ind.* **12**, 49.

REVIEWS

Demand Analysis. A Study in Econometrics. By HERMAN WOLD, in association with LARS JURÉEN. New York: Wiley and Sons (Stockholm: Almqvist and Wiksell). Pp. 1953. xvi + 358. 56*s*.

Originating in a study of consumer demand for food in Sweden undertaken in 1938 by Prof. Wold and Mr Jureen, this book sets out to present a self-contained account of both methods and results. It includes a report on the empirical findings, based on family budget surveys and market statistics, a section giving a survey of methods in non-technical language, and sections dealing in more concentrated and technical manner with some aspects of economic theory, stochastic processes and regression analysis.

The theoretical sections are of considerable incidental interest for econometrics generally, although the developments, on an abstract level, are occasionally only remotely relevant to the reported empirical study.

The author states that one of his principal aims was to justify the use of the traditional method of least-squares regression, via the use of economic models of so-called recursive type. His main argument rests on the notion of causality in consumer demand situations. With this as a guide to the method of application, he obtains, among others, the result that the regression method for the estimation of recursive systems is unbiased and consistent on the assumption of non-correlation between the explanatory variables and the disturbances, irrespective of whether there is inter-correlation between these disturbances. A generalization of the classical formula is derived for the standard error of a regression coefficient which allows for intercorrelation.

The wide scope of recursive models is indicated by a proof that any set of time series has a formal representation in terms of a recursive system.

However the causality argument is regarded (and it is certainly not a straightforward one), the regression method would still appear on Wold's showing to compare favourably, in demand analysis at least, with other methods which have been proposed.

The book is very fully annotated; there is an extensive bibliography (although several of the dated references in the text are omitted), and important sets of exercises accompany the theoretical parts. Based on years of research into the foundations of the subject, this is a very thorough work of scholarship.

F. G. FOSTER

A Study in the Analysis of Stationary Time-series. 2nd edition. By HERMAN WOLD, with an Appendix by PETER WHITTLE. Stockholm: Almqvist and Wiksell. 1954. Pp. viii + 236. 42*s*.

As far as the main text is concerned, the present edition differs little from the first, published in 1938, the only change of major importance being the replacement of appendices A and B in the first edition by two new ones.

In the first appendix, Prof. Wold makes short but useful comments on certain parts of the text, mainly in the light of recent work. Appendix 2, comprising 32 pages, is written by Dr P. Whittle and is devoted to a survey of recent advances in the subject.

The first three chapters of the book dealing with Prof. Wold's fundamental contributions to the theory of discrete stationary schemes are by now familiar to most statisticians and need no further comment. Chapter IV, on the application of stationary schemes to experimental data, has not been as successful in withstanding the passage of time, and the discussion tends to be outmoded in view of the advances made in the testing of specific hypotheses and in the problem of estimation.

Chapter 1 of Appendix 2 is devoted to the contributions of Cramér, Kolmogoroff and Wiener to spectral theory and Chapter 2 to a synopsis of Whittle's own work in the theory of inference, together with some miscellaneous results in distribution theory and periodogram analysis. To attempt even a cursory review of the field in so short a space is not practicable, and the result is a one-sided exposition of the subject. Even so, the treatment is interesting and some new results are introduced.

The emphasis laid throughout on the mathematical and structural, rather than the statistical aspects of time-series analysis, is likely to make the book of greater utility to the mathematical statistician than to one who is interested in the applications to economics, physics, meteorology, etc. The latter is also likely to view with extreme pessimism the warnings made at intervals throughout the book about the reliability of the information provided by the analysis. Briefly stated, the argument is that significance problems are complicated by the fact that one is determining the whole structure of a series from a realization or sample and also that the attachment of quantitative significance to the sample functions is not possible since they are conditioned by the size of the statistical masses to which they refer; in particular, there is arbitrariness in (*a*) the time unit of collection of the data, (*b*) the region in space to which this unit refers. A great deal has been done to remove the first objection by realizing the futility of testing individual serial correlations, and much progress has been made in discriminating between various hypotheses. The evidence up to the present seems to indicate that the problem is soluble in terms of the classical theory.

The second objection is more troublesome but is not so serious as one might be led to expect. As far as (*a*) is concerned, the choice of time unit raises no special difficulty, since in many instances the autocorrelation properties of time-series are not in themselves of fundamental importance. Thus whilst collecting data at quarterly intervals instead of annually is likely to alter the structural form of the series, if it were required to estimate the correlation between two time-series, the ultimate aim would be to eliminate the effect of autocorrelation during the course of the analysis. The intrinsic autocorrelative properties of a series are likely to be more important in the problem of prediction, but recent work seems to suggest that as far as practical applications are concerned much remains to be done in this field.

The difficulties raised by (*b*) are more serious, but this type of objection is latent in, and tends to restrict the conclusions which are drawn in many other problems in statistics.

The need for further research is obvious, and Prof. Wold's own work on this problem of quantitative significance is sufficient to raise the question that perhaps our assumptions in dealing with observed series are too stringent.

G. M. JENKINS

An Introduction to Stochastic Processes with special reference to Methods and Applications. By M. S. Bartlett. Cambridge University Press. 1955. Pp. xiv + 312. 35*s*.

Probability problems concerned with the description of changes with time—where we use 'time' in the operational sense—are as old as the theory itself. The entire development of the subject in the sixteenth and seventeenth centuries was due to the zeal and passion of gamblers and, among other queries, the probability of ruin was one which ranked importantly. Interest in this type of problem never faded out, and we find it discussed by one after another of the great probabilists right up to the present day. The difference between the moderns and their predecessors is that problems involving the dynamics of change are now given a special title—stochastic processes—and are subject to a unified systematic attack.

This systematic attack, which has its origins chiefly in the research of the Russian school of probability and in particular in that of Kolmogoroff, has resulted in a great deal of elegant mathematics and a certain amount of useful statistical methods. Its great defect has been that the problem of stochastic processes has been lifted beyond the understanding of many statisticians who still continue to solve the problems in *ad hoc* fashion as they arise.

In 1946 M. S. Bartlett attempted to remedy this state of affairs by publishing in mimeographed form the lectures which he had given in the University of North Carolina. Further clarification, in the discrete case, came with W. Feller's book *Introduction to Probability Theory*, and now we again have Prof. Bartlett who attempts to avoid too much mathematical abstraction and gives a general survey of techniques and applications. Topics on pure theory covered by him are random sequences, processes in continuous time, limiting stochastic operations and stationary processes. The problem of the random walk, Markoff chains, renewal processes, stochastic convergence and spectral analysis are included. Under statistical applications we have a chapter on applications of the random walk, queues, population growth and epidemic models, a chapter on statistical inference in stochastic processes and a chapter on the correlation analysis of time series including harmonic analysis.

This book is the best which has yet been produced on this specialized topic. It is lucidly written by someone who clearly sees the statistical implications of the abstract theory and may be read without undue difficulty by any student who has two years of university mathematical training behind him. It will undoubtedly be found indispensable by anyone wishing to learn something of a subject which has engaged the attention of nearly all probabilists in the past decade.

F. N. DAVID

Probability Theory. By M. Loève. New York: D. van Nostrand Company, Inc. 1955. Pp. xv+501. $12.00.

This is a book on probability theory written by a mathematician who has himself made distinguished contributions to the subject during the past twenty years. Prof. Loève divides his book into five parts. In the introductory part we have what might be called the intuitive approach to probability which is very often all that the statistician needs. In spite of the fact that the topics are not new they are set out in clear and pleasing fashion, and the corollaries to the various theorems proved, particularly in the laws of large numbers, will be fresh to those who are not thoroughly familiar with the French and Russian literature. The notion of chain dependence is also introduced at this stage. All this is introductory.

In Part One proper Prof. Loève sets us sternly to work with 83 pages on 'Notions of Measure Theory'. Here we have additive set functions, topological spaces, measurable functions and Lesbesgue integration. The two chapters which compose this part can be read fairly easily by a student who has taken a university degree course in mathematics, but the reasoning is rather condensed and may cause difficulty to students coming to it for the first time. In Part II, entitled 'General Concepts and Tools of Probability Theory', we have the development of probability laws, characteristic functions and distribution functions, much of which is in essence familiar although not perhaps the rigorous presentation.

With 'Independence', which is the topic of Part Three, Prof. Loève gives us a characteristically thorough thrashing out of convergence in probability and central limit theorems. Little of this has appeared explicitly in English, and it is useful to have it all down in a moderately compact form (107 pages). From 'Independence' it is natural that we should turn to 'Dependence', the topic of Part Four. Here, apart from the clear and mathematically rigorous presentation, there is little which has not been covered by Doob and Bartlett, except perhaps the second-order random functions.

This book will be excellent for those with a mathematical training who wish to specialize in probability theory regarded as a mathematical discipline. It is throughout logically inevitable in its presentation and written as rigorously as only a French-trained mathematician can. There can be no better beginning for a young mathematical student than to be given this book. It will, however, not be helpful in the training of the statistics student. Many a promising statistician has been wrecked on the desert island of mathematical rigour, and those trained in modern mathematical techniques often find the necessarily heuristic approach to numerical data painful. However, for the class of student indicated the book is excellent and all providing. It is unfortunate that it has to be recorded that the price is $12.

F. N. DAVID

Statistical Inference. By Helen M. Walker and Joseph Lev. New York: Henry Holt and Company. 1953. Pp. xi+510. 48*s*.

This text-book on statistical method is designed for beginners with little or no mathematical background. A good many mathematical formulae are, however, used (albeit always accompanied by careful explanation and practice exercises), and a student who has mastered such a text as Prof. Walker's previous book, *Mathematics Essential for Elementary Statistics*, would find the going easier. In choice of subject-matter and general emphasis it is definitely 'applied', the application being to educational research; this is not, however, revealed by the title.

In scope it is fairly comprehensive. It begins with the general idea of statistical inference and an intuitive discussion of hypothesis testing. The concept of probability is introduced, then the binomial distribution, with a discussion of the closeness of the normal approximation. By use of a population of two classes only, a variety of problems of hypothesis testing and estimation are treated, the account including such topics as one- and two-sided tests, critical regions, errors of the first and second types, power functions, confidence limits—concepts which gain in clarity for the beginner by definition in this simple context. Populations with more than two classes are next studied, and then some of the more important concepts regarding distributions on a continuous variable. There follows an account of a class project in sampling which is designed to familiarize the reader with the statistics possessing one of the four continuous distributions of major importance: the Normal, Chi-square, Student's and F distributions. The method here is to plot the empirically derived distribution, superimpose the mathematical curve, and then introduce the student to the table of this function in place of its mathematical formula. With this basic material assembled, the subjects now treated include: inferences concerning

the mean or difference between two means; inferences concerning variances and standard deviations; analysis of variance; linear regression and correlation; biserial, point biserial, phi, tetrachoric and rank-order correlation coefficients and the correlation ratio; the effects of measurement errors and the reliability coefficient; multiple regression and correlation; analysis of variance with two or more variables of classification; analysis of covariance; percentiles; transformation of scales; non-parametric methods. The chapter on non-parametric methods was written by Prof. Lincoln Moses.

The treatment is concrete, and as far as possible basic ideas are explained by means of numerical examples. The student is drilled in computing techniques; a rather extensive set of twenty-three tables is an integral part of the text for this purpose, detailed instructions in their use being given. (The reader may on occasion be somewhat startled at being addressed directly from the text, exhorted to do such and such a computation.) A welcome feature of the book for educationalists should be the amount of illustrative numerical data actually taken from published work on educational research.

The book is well referenced, and a list of authors for further reading is placed at the ends of chapters. Many exercises are set, with answers given at the back of the book. There is a subject index, an author index and a glossary of the mathematical symbols used.

It is to be expected of a text with which Prof. Walker is associated that the exposition should be pedagogically sound, and there is evidence that a great deal of care was taken about the order and method of presentation of the material. The result cannot be said to be completely satisfactory. It is, for example, open to serious doubt on occasion whether the average student will be able to absorb simultaneously details of computing techniques and the basic principle of the statistical method under explanation. Occasionally, too, the style becomes so woolly as to present quite inaccurate information. Thus, on p. 14, two observations are defined to be independent 'when information about one of them provides no clue whatever as to the other'. A reader may well be puzzled by this if he considers that in the case of two independent observations from the same (unknown) population, the first observation provides an unbiased estimate of the mean of the second.

Despite some defects, however, research students of education and allied fields should find this text invaluable, not only as an instructional manual, but also as a general work of reference in statistical method.

F. G. FOSTER

Outline of Biometry and Statistics.. By C. I. Bliss and D. W. Calhoun. New Haven, Conn.: Yale Co-operative Corporation. 1954. Pp. 272 + xvi. $4.50.

This book will be extremely useful to those who teach biometry, and to statisticians who work in a biological field or who wish to have biological examples to illustrate their teaching of statistics. Each chapter consists of a number of very concise explanatory notes, together with several illustrative worked examples and exercises chosen from a wide range of experimental material.

The subjects covered include the binomial, negative binomial and Poisson distributions; chi-squared and the analysis of contingency tables; the normal distribution with the estimation of its mean and variance; the analysis of variance, factorial experiments and simple experimental designs; regression, correlation and association; bio-assay. The authors regret that the proposed length of course (about 200 hours of students' time) necessarily excludes a few other topics such as covariance, probits, discriminants, and sequential and other sampling methods.

The professional statistician will find this book packed with information, such as short-cut formulas, precise statements of the limits within which approximate tests can be relied upon, and alternative parametric and non-parametric tests. A number of useful tables and charts are included. A long list of references enables the reader to consult the original papers whenever necessary, except for work yet unpublished. The exposition is almost always extremely lucid, and the reviewer has noticed few errors (apart from those already covered by a list of errata). But surely the exact test for 2×2 tables was independently discovered by Fisher and by Irwin (see *Metron*, **12**, 1935, 1)?

The book is too condensed for the unassisted student, but one working through the course under a teacher's guidance will find it very useful.

CEDRIC A. B. SMITH

Statistics and Mathematics in Biology. Edited by O. KEMPTHORNE, T. A. BANCROFT, J. W. GOWEN and J. L. LUSH. Ames, Iowa: Iowa State College Press. 1954. Pp. ix + 632. $6.75.

In case the title should mislead, it should be made clear at once that this is not a text-book on the use of statistical and mathematical methods in biology. It is a collection of forty-four essays by various authors (many of great distinction) on a wide range of topics in which biology and statistics meet. In many cases these deal with recent research by the author; some are highly theoretical, others concern practical problems involving only simple mathematics. Almost all are lucidly expressed, and can be read with interest and profit. The subjects discussed include causation, regression, path coefficients and multivariate analysis, classification problems, experimental designs, competition between species, growth curves, sampling of populations, toxicity tests, bio-assay, taste testing, feeding experiments, animal behaviour, gene frequency estimation, the effect of radiation on cells and viruses, genetic and breeding problems and the estimation of nucleoproteins in cell division.

Naturally this is not the sort of book to which one can turn for complete and detailed information on any one topic, although the very full bibliography may indicate the source of such information. But it is one which it would be useful to have about the laboratory (whether statistical or biological), since somewhere it may suggest the appropriate line of approach to a particular problem, or indicate a profitable new line of research.

One minor point may be mentioned. K. R. Nair has pointed out (*Bull. Calcutta Statist. Ass.* **20**, 1954, 18) that Quenouille's 'almost balanced incomplete block designs' are particular cases of the Bose-Nair-Rao 'partially balanced incomplete blocks' (Bose & Nair, *Sankhyā*, **4**, 1939, 337).

CEDRIC A. B. SMITH

Statistical Analysis in Chemistry and the Chemical Industry. By C. A. BENNETT and N. L. FRANKLIN. New York: John Wiley and Sons, Inc. 1954. Pp. xvi + 724. 58*s*.

This book contains a well-written account of statistical techniques, some of which are not readily available elsewhere. After an introductory chapter, general notions of frequency distributions, measures of location, spread and association are introduced. There follows a discussion of probability theory including some account of moment generating functions, cumulants and k statistics and the common sampling distributions derived from the normal law. A short account of the properties of gamma and beta functions is followed by a chapter on confidence limits and tests of significance, including some non-parametric methods and a somewhat brief discussion of sequential tests. In Chapter 6 there is an account of linear regression, curvilinear regression and discriminant functions and of methods for the solution of the linear equations which arise in the application of these techniques.

Chapter 7, which contains over 150 pages, is concerned with the analysis of variance, and this is followed by a discussion of almost equal length on the design of experiments. The book ends with chapters on the analysis of counted data, quality control and tests for randomness.

To those workers in the chemical industry and elsewhere who have been introduced to statistical methods via a 'cookery book' and who now wish to broaden their knowledge and to learn something more of the basis of these methods, this book is recommended. It will be less useful to those who wish to know how statistics should be used in practice. Some of the examples fail to demonstrate the value of the method discussed, or the sort of circumstances in which it might be used, and instances occur where the statistical analysis confuses rather than clarifies the situation. This is particularly true of the example on p. 395, where the situation is much more readily appreciated from a set of four simple graphs than from the rather complicated analysis of variance which is given in this book.

It is very important that the experimenter should be the master and not the slave of statistical methods. In particular, he must be very clear about what it is he really wants to know and should not allow himself to be conditioned into using statistical ideas and concepts which are not really appropriate to his investigation. In particular, he should not allow himself to be overawed by apparently complicated mathematical machinery. When he feels his problem is not really answered by the application of standard techniques he should try to deal with it from first principles. His solution even if it is somewhat approximate and unorthodox will usually serve him better than the misapplication of a mathematically 'exact' technique. In helping to instil these first principles this text-book is valuable; the reader should, however, not take too seriously some of the applications which are described.

G. E. P. BOX

A Million Random Digits with 100,000 Normal Deviates. The RAND CORPORATION. Illinois, U.S.A.: The Free Press. 1955. Pp. xxv + 600. $10.00.

Since L. H. C. Tippett's set of 40,000 random numbers was first published in 1927 as *Tracts for Computers*, No. xv, there has been a steady demand for random numbers from a wide variety of users. This was recognized by the publication of M. G. Kendall's and B. Babington Smith's set of 100,000 digits in *Tracts for Computers*, No. xxiv, just prior to the war. Since then the demand has grown rather than diminished, and the recent emphasis on Monte Carlo methods has led to the need for an even larger supply. This need has now been filled by the publication of this volume of 1,000,000 random digits produced with the aid of an electronic roulette wheel by the Rand Corporation. Tests for randomness, such as the frequency test, the so-called poker test and the pairs test, have been applied to the numbers and give satisfactory results.

Half the table has been used to obtain 100,000 random normal deviates by the conversion of five-digit random numbers with a table of the cumulative normal distribution function. The deviates obtained are tabulated to three decimal places.

P. G. MOORE

Tablitsi dlya vichisleniya nepolnoi Γ-funktsii veroyatnosti χ^2 (Tables for the calculation of the incomplete Γ-function and the χ^2-probability function). By E. E. SLUTSKII (Ed. A. N. KOLMOGOROV). Moscow and Leningrad: Izdatelstvo Akademii Nauk SSSR. 1950. Pp. 14 + 55 pp. tables.

These tables are intended to facilitate the computation of

$$P(\chi^2, n) = \frac{1}{2^{\frac{1}{2}n-1}\Gamma(\frac{1}{2}n)} \int_{\chi}^{\infty} x^{n-1} e^{-\frac{1}{2}x^2} dx,$$

and the related incomplete gamma-function. There is a table of $P(\chi^2,n)$ for

χ^2: 0 (0·1) 3·2 and χ^2: 3·2 (0·2) 7·0 (0·5) 10·0 (1·0) 35·0,
n: 0 (0·05) 0·2 (0·1) 6·0 n: 0 (0·1) 0·4 (0·2) 6·0.

There are also tables of three auxiliary functions designed for special purposes. To facilitate interpolation for small values of χ^2 the function

$$T(\chi^2, n) = (\tfrac{1}{2}\chi^2)^{-\frac{1}{2}n}(1 - P(\chi^2, n))$$

is tabled for

χ^2: 0 (0·05) 0·2 (0·1) 1·0,
n: 0 (0·05) 0·2 (0·1) 6·0.

For larger values of n the function $P(\chi^2, n)$ is tabled, with argument

$$t = \sqrt{(2\chi^2)} - \sqrt{(2n)}$$

in place of χ^2, for

t: −4·0 (0·1) 4·8,
n: 6·0 (0·5) 11·0 (1·0) 32·0.

A further table gives values of the function $\pi(t, x) = P(\chi^2, n)$, where t is as above and $x = \sqrt{(2/n)}$. This function is tabled for

t = −4·5 (0·1) 4·8,
x = 0 (0·02) 0·22 (0·01) 0·25.

It is claimed that this small (seven page) table effectively provides for the computation of the incomplete gamma-function in the region ($n > 102$) not covered by the existing tables of K. Pearson.

A final table contains coefficients of Everett's and Newton's interpolation formulae. Second and fourth central differences with respect to χ^2 (or t), and second central differences with respect to n (or x) are printed in the tables.

The tables are clearly reproduced and there is an introduction setting out their genesis, calculation and use, with a number of worked examples.

N. L. JOHNSON

OTHER BOOKS RECEIVED

1. **Rank Correlation Methods**, 2nd edition. By M. G. KENDALL. London: Charles Griffin and Co. 1955. Pp. 196. 36*s*.
2. **Manpower Shortage and the Fall of the Roman Empire.** By A. E. R. BOAK. U.S.A.: University of Michigan Press. 1955. Pp. 169. $4.50.
3. **The Foundations of Statistics.** By L. J. SAVAGE. U.S.A.: John Wiley and Sons; London: Chapman and Hall. 1954. Pp. xv + 294. 48*s*.
4. **Statistics in Research.** By B. OSTLE. U.S.A.: Iowa State College. 1954. Pp. xiv + 487. $6.95.
5. **Annual Epidemiological and Vital Statistics for 1952.** Switzerland: World Health Organization, Geneva. 1955. Pp. x + 533. 50*s*.

Publications of the U.S. Department of Commerce, National Bureau of Standards

6. **Tables of Sine and Cosine Integrals for arguments from 10 to 100.** Applied Mathematics Series, 32. 1954. Pp. xv + 187. $2.25.
7. **Tables of the Gamma Function for Complex Argument.** Applied Mathematics Series, 34. 1954. Pp. xvi + 105. $2.00.
8. **Tables of Functions and of Zeros of Functions.** Applied Mathematics Series, 37. 1954. Pp. ix + 211. $2.25.
9. **Contributions to the Solution of Systems of Linear Equations and the Determination of Eigen-values.** Edited by OLGA TAUSSKEY. Applied Mathematics Series, 39. 1954. Pp. 139. $2.00.
10. **Tables of the Error Function and its derivatives.** Applied Mathematics Series, 41. 1954. Pp. xi + 302. $3.25.